Single-Neuron Sequencing to Explore Somatic Genetic Variants in Normal and Pathological Human Brain Development

Single-Neuron Sequencing to Explore Somatic Genetic Variants in Normal and Pathological Human Brain Development

A dissertation presented

by

Xuyu Cai

to

The Division of Medical Sciences

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

February, 2013

Dissertation advisor: Dr. Christopher A. Walsh Author: Xuyu Cai

Single-Neuron Sequencing to Explore Somatic Genetic Variants in Normal and Pathological Human Brain Development

Abstract

The human brain is one of the most exquisite structures in nature, featuring extreme functional complexity and capacities that allow for advanced cognitive abilities. During the development of the human brain, neural progenitors undergo massive proliferation, which is known to inevitably result in spontaneous mutations; yet the degree of somatic mosaicism within the human brain is unexplored. Several hypotheses have been proposed that various types of somatic mosaicism may serve as an adaptive mechanism to diversify neurons and thereby promote the functional complexity of human brains. Previously proposed mechanisms to increase somatic mosaicism within the brain include elevated somatic LINE-1 element retrotransposition, and the creation of somatic aneuploidy during neurogenesis. On the other hand, genomic diversity needs to be balanced by genomic stability, in order to protect against deleterious mutations that reduce the fitness of the cells, or oncogenic mutations that might promote cancers. In fact, brain-specific somatic mutations have also been proposed to contribute to the unexplained burden of neurological diseases. To directly study genomic variability from cell-to-cell within the human brain, we developed a method to isolate and amplify single neuronal genomes from postmortem and surgically resected human brain tissues. We quantified the frequency of somatic LINE-1 retrotransposition events and aneuploidy in

human cortical neurons, and found that the frequencies of both are low, with no sign of brain-specific elevation, arguing against the hypotheses that these two mutational sources are obligate generators of neuronal diversity. Additionally, aneuploidy analysis was performed on bulk and single cortical cells from a hemimegalencephaly brain. Hemimegalencephaly is an asymmetrical brain overgrowth syndrome caused by somatic mutations in brain. Single-cell analysis identified an unexpected mosaic tetrasomy of chromosome 1q, affecting both neuronal and glial populations, as a genetic cause of hemimegalencephaly. These results demonstrate that single-neuron sequencing allows systematic assessment of genomic diversity in the human brain and the identification and characterization of pathogenic somatic mutations underlying neurological disorders.

Table of Contents

Acknowledgements

I'd first like to thank everyone from the Walsh lab, especially my PhD mentor, Dr. Chris Walsh. He gave me tremendous flexibility and support in pursuing my scientific dreams in his lab. Without his consistent encouragement along the way, our project would not have been possible. He is also a great scientist with the big vision, full of great ideas and insights that guided me though my PhD. He is also a great mentor outside of science, full of compassion and always willing to listen and help on all aspects of my life. Gilad Evrony, my coworker on the single-cell project, is the second person I want to thank from the Walsh lab. Gilad is one of the most brilliant people I've ever met. The most enjoyable part of my entire PhD were the last night brainstorming sessions with Gilad, on our project as well as on science in general. He is not only full of great ideas; he is also serious about the execution of his ideas with full passion. It has been my great good fortune to work with Gilad during almost my entire PhD, to learn from him and to grow together into mature scientists. I would also like to thank my bay mates, Divya and Byoung-Il, who always kept me company, through numerous late nights. Wen, my workout buddy, forced me into the gym to get my positive life attitude back after having a baby. She also encouraged me to run a marathon, which I'm currently working on. Xiaochang, the only other Chinese in the lab, not only helped me keep connected with my culture, but also taught me how to genotype mice and how to be a parent. I'd also like to thank Christina Elhosary, Ben Hills, Bhaven Mehta, Hillel Lehmann and Deniz Kutaby for their help on our project. Christina has always been a great friend, who not only

helped me to run thousands of PCRs, but also spent lots of time outside of the lab with me, to drag me up from the lowest point of my life.

I would also like to thank my family. My parents, especially my mother, Grace, has offered me tremendous help throughout my PhD. I simply cannot imagine how I could possibly have gotten to this point without her help in taking care of Liyan for me, and providing mental support through the tough time. She is also a fighter herself as she has been fighting with breast cancer for the past 3 years and staying healthy. I'd like to thank my son, Liyan, as well. He taught me how life is truly a miracle and how lucky I am to have all my family surrounding me.

I'd like to thank my dissertation advisory committee, Jeff Macklis (chair), David Pellman, and Roz Segal, as well as my dissertation examination committee, Jeff Macklis (chair), Paul Blainey, Matt Anderson and Gabriel Corfas, for taking the time to guide me through this unprecedented 5 years of my life, helping me to achieve my scientific goals.

Chapter 1: Introduction

Overview of Current Knowledge on Human Brain Development and Unanswered Questions

Evolution of human brain

The evolution of the human brain somehow resulted in the exceptional cognitive abilities that make us human. It is generally believed that the enormous expansion in human brain size compared to other closely related primate species directly enabled the exquisite complexity and diversity of neural cell types and connections that characterize the human brain (Rakic, 2009; Lui et al., 2011). Historically, studies on human brain development have been mainly descriptive at the anatomical and histological level. However, recent technological advances finally allow studies of human brain-specific features at the molecular and cellular level. These studies have revealed human-specific gene compositions, transcriptional networks and cell types, and have further elucidated the uniqueness of the human brain, yet also raise challenges to assessing the functional significance of these distinctive features in contributing to human cognitive capacities (Konopka et al., 2009; Hawrylycz et al., 2012; Konopka et al., 2012; Zeng et al., 2012). Notably, the prefrontal cortex was identified to have the most prominent human-specific expression profile as opposed to other brain regions (Konopka et al., 2012), consistent with its primary role in higher cognitive functions and its prominent expansion in surface area relative to total brain volume during human brain evolution (Rakic, 2009).

Neocortical development

Human cerebral cortex has been extensively characterized based on gross anatomy, cytoarchitecture, and topography. Neocortex, defined by its six-layer

cytoarchitecture, is the most recently evolved cortical region, is distinctive in mammals, and is the center of higher cognitive functions (Rakic, 2009). Compared to other brain regions, the structure of the neocortex is relatively uniform. It is organized into a six-layer structure, specified by neuronal cell types and connections (Rakic, 2009). To date, hundreds of neuronal subtypes have been identified in the neocortex, based on their morphological and molecular features, as well as their projection patterns. These neuronal subtypes can be categorized into two major classes: excitatory pyramidal neurons derived from the dorsal ventricular epithelium, and inhibitory interneurons derived from the ventral and possibly also the dorsal ventricular epithelium (Molyneaux et al., 2007; Zecevic et al., 2011). The distinctive origin of the two major neuronal cell types reflects an important rule of cortical development, such that brain cells derived from distant cell lineages tend to have different functions. However, how much these distinct cell lineage relationships contribute to the generation of the enormously diversified cortical neuronal subtypes remains largely unknown (Molyneaux et al., 2007).

The regionalization of the neocortex is initiated by patterning centers that secrete signaling factors such as FGFs, WNT and BMP at the neural tube stage before the onset of neurogenesis (Kiecker and Lumsden, 2012). The maintenance and fate specification of neural progenitor cells are further regulated by a series of cell autonomous and non-autonomous mechanisms during neurogenesis. Based on current knowledge, there are two major neurogenic germinal zones: the ventricular zone (VZ), which is conserved from primates through rodents; and the outer subventricular zone (OSVZ), which is much more prominent in gyrencephalic animals (Breunig et al., 2011; Lui et al., 2011) (**Figure 1-1**). It has been hypothesized that the massive expansion of the OSVZ is responsible for

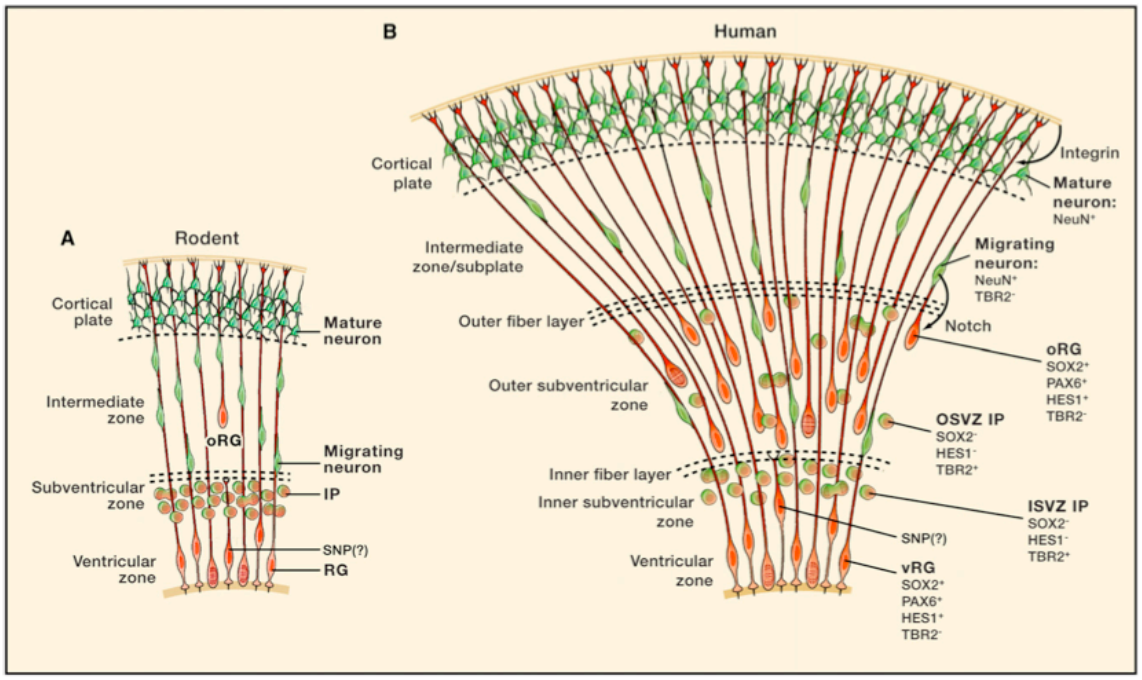


Figure 1-1. Expansion of OSVZ progenitors in human compared to rodents. Adapted from Lui et al. (2011)
(A) Current view of rodent corticogenesis. The germinal zone is made up by the ventricular zone (VZ), populated by radial glial cells (RG) and subventricular zone (SVZ), populated by intermediate progenitors (IP). Very few outer radial glia (oRG) are observed from rodent brains.
(B) Current view of human corticogenesis. In addition to the VZ and SVZ, a prominent outer subventricular zone (OSVZ) populated by oRG and OSVZ IP cells are observed in developing human cortex. The expansion of OSVZ population is thought to account for the tangential expansion of human cortex surface area.

the elaboration of cortical gyri, and the tangential expansion of neocortical surface area in primates and human (Hansen et al., 2010; Lui et al., 2011; Reillo et al., 2011), but this remains unproven. Both the VZ and OSVZ contain heterogeneous progenitor cell populations, which can be broadly classified into self-renewing radial glia (RG and oRG) and transit-amplifying neurogenic intermediate progenitors (IP). Understanding of the cell lineage relationship between different OSVZ progenitor cell types and their extremely diversified neuronal progenies remains a major challenge in understanding human cortical development and evolution.

Open questions and challenges

The recent discoveries on OSVZ progenitors, human specific neuronal subtypes and human specific patterns of gene expression have begun to shed some light on the cellular and molecular mechanism of human brain evolution (Lui et al., 2011; Hawrylycz et al., 2012; Konopka et al., 2012; Zeng et al., 2012). It is conceivable that new classes of neurons allow the formation of new patterns of connections, which increases the capacity and flexibility of cognitive abilities. It is tempting to ask whether the expansion of OSVZ progenitor pool in human brain contribute to its increased neuronal diversity, and whether there exists human specific progenitor cell populations that give rise to particular neuronal subtypes enabling our higher cognitive functions. Development of new cell lineage tracing techniques in human brain will help to start addressing these questions.

Brain malformations as an avenue to study brain development

Our increasing knowledge of human brain development helps us to understand the disease mechanisms of common neurodevelopmental disorders, such as autism spectrum disorders (ASD), schizophrenia (SCZ), intellectual disability (ID) and epilepsy (LaMonica et al. 2012). Meanwhile, human neurodevelopmental disorders provide great opportunities to study normal brain development at the molecular level. Brain malformations are rare and severe forms of neurodevelopmental disorders with anatomical alteration of brain structures that are easily identified radiographically. Brain malformations are often caused by the disruption of key aspects of brain development, such as neurogenesis, neuronal migration, and axon guidance. Therefore, brain malformations are often associated with one or multiple neurological conditions such as ID, ASD and epileptic seizures. On the other hand, more common and milder neurological and neuropsychiatric conditions, such as non-syndromic ASD, non-lesional epilepsy (epilepsy conditions without radiographically identifiable lesions), SCZ and bipolar disorder (BIP), are associated with more subtle defects of neurological function, such as synaptic function and ion channels (Sullivan et al. 2012; State & Sestan 2012). Therefore, studies of severe brain malformation disorders often lead to understanding of the more fundamental aspects of brain development. For instance, genetic disorders resulting in abnormal brain size provide tremendous insights into the regulation of neurogenesis that ultimately determines brain size (Gilmore & Walsh 2012).

Microcephaly

Human autosomal recessive primary microcephaly (meaning "small brain") (MCPH) is a genetically heterogeneous neurodevelopmental disorder defined by marked reduction in brain size with grossly preserved brain architecture at birth. Clinical features of MCPH also include

intellectual disability with variable severity and occasional seizures (Thornton & Woods 2009). To date, more than ten genes have been identified to cause this rare genetic disorder; and remarkably, most of these genes encode proteins associated with centrosomal and microtubule-related cellular functions, with a few exceptions that are involved in DNA damage repair (Gilmore & Walsh 2012; Shen et al. 2010). The centrosome plays multiple roles in regulating cell divisions and cytoskeletal reorganization (Bettencourt-Dias et al. 2011) (**Figure 1-2**). The functional convergence between centrosome and DNA damage repair is cell cycle regulation, which empirically confirms the speculation that brain size is largely regulated by proliferation of neural progenitor cells during neurogenesis. It is interesting to note that despite their ubiquitous expression and expected cellular function throughout the body, most MCPH-associated centrosomal mutations cause brain-specific defects, with a few exceptions such as *PCNT*, and some cases in *CEP152* and *CENPJ* (Al-Dosari et al. 2010; Kalay et al. 2011; Griffith et al. 2008), suggesting that brain is more susceptible to cell proliferation defects. This is presumably due to the extremely high demand in progenitor pool expansion within the small timeframe of neurogenesis.

Lissencephaly

Lissencephaly (meaning "smooth brain") is another genetic neurodevelopmental disorder characterized by simplified gyrification that leads to a smooth cerebral surface, mental retardation and seizures, as a consequence of severe defects in neuronal migration (Wynshaw-Boris et al. 2010). Mutations of a number of genes have been identified as the genetic causes of lissencephaly, including *LIS1*, *DCX*, *TUBA1A* and *RELN*, all of which are involved in cytoskeletal organization and thereby are essential for proper neuronal migration (Wynshaw-

Boris et al. 2010; Manzini & Walsh 2011). Although it was originally thought that microcephaly and lissencephaly are separate disorders with distinct cellular processes being affected during neurodevelopment, the identification of *WDR62* mutations causing both profound microcephaly and simplified gyrification blurs the boundary of the two disorders and suggests cellular pathways that intersect neural progenitor proliferation and neuronal migrations (Nicholas et al. 2010; Yu et al. 2010; Bilguvar et al. 2010). Most lissencephaly cases are caused by mutations involving *LIS1* or *DCX*. Human *LIS1* mutations are mostly *de novo* dominant mutations, involving the heterozygous deletion of 17p13.3, which leads to haploinsufficiency of the *LIS1* gene. *DCX* mutations are the major genetic causes of X-linked lissencephaly. Hemizygous males show classic lissencephaly phenotype, whereas heterozygous females show mosaic phenotype of an extra subcortical band, referred as "double cortex syndrome" (Walsh & Engle 2010; Gleeson et al. 1998).

Hemimegalencephaly

Hemimegalencephaly (HME) is a sporadic epileptic disorder characterized by the enlargement and malformation of one cerebral hemisphere (Flores-Sarnat et al. 2003). Two emerging studies have pinpointed somatic hyperactivation of the PI3K-AKT-mTOR pathway to be a major genetic cause of this disorder (Poduri et al. 2012; Lee et al. 2012). PI3K-AKT-mTOR pathway has been extensively characterized as a signaling pathway that responds to external growth factors to regulate cell proliferation, growth and survival; hyperactivation of the pathway is frequently observed in human cancers (Engelman et al. 2006) (**Figure 1-3**).

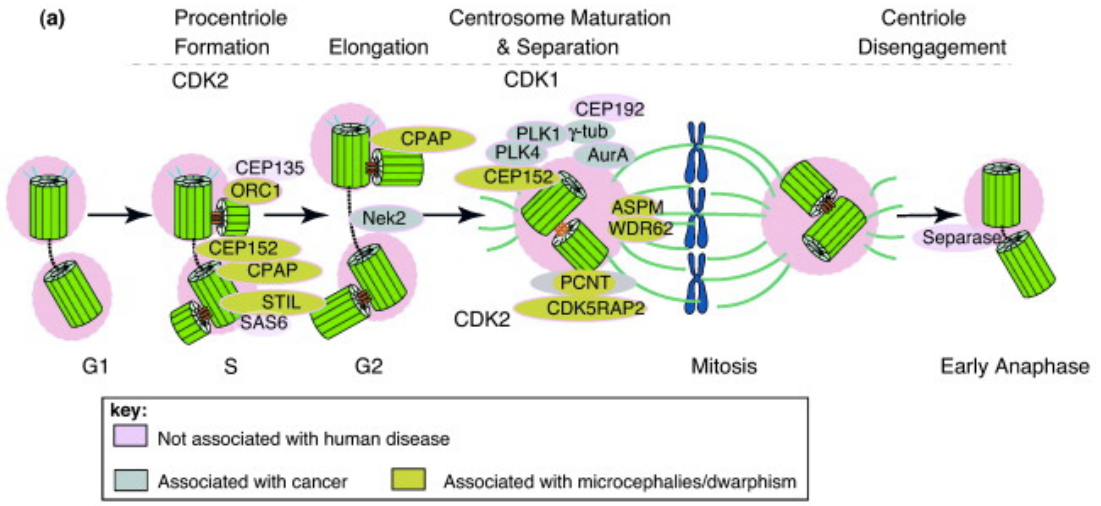


Figure 1-2 Summary of centrosomal genes involved in microcephaly. **Adapted from Bettencourt-Dias et al. (2011).**
Mutation involving multiple aspects of centrosome biology including procentriole duplication, elongation and centrosome maturation/separation were all identified, suggesting the convergence of a molecular pathway in causing a particular disorder.

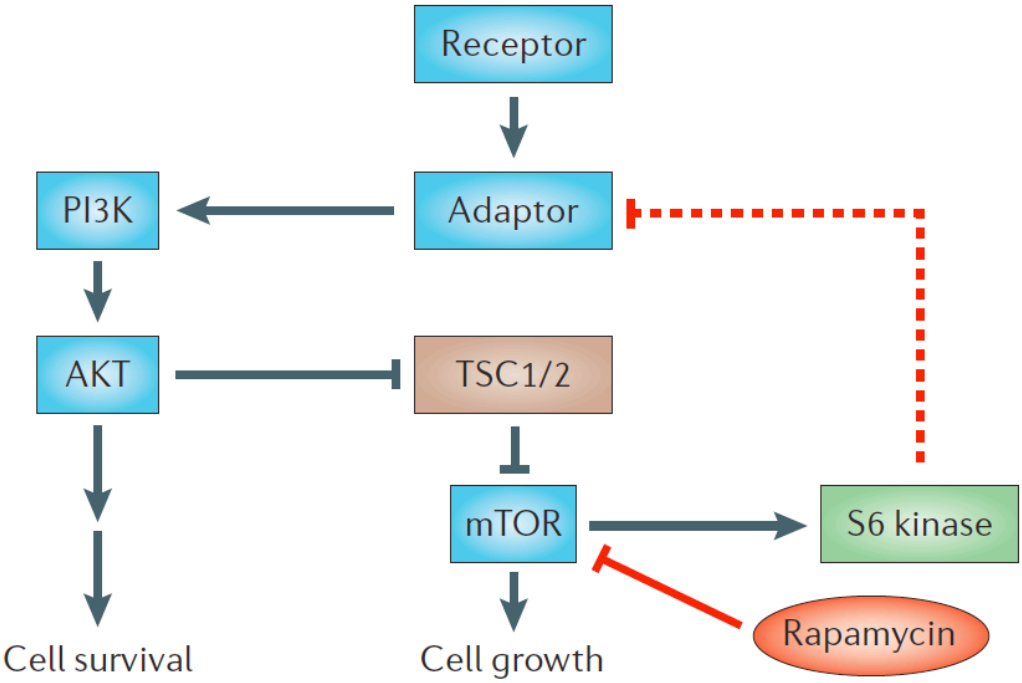


Figure 1-3. PI3K/AKT/mTOR molecular pathway regulates cell growth, survival and proliferation. Adapted from Engelman et al. (2006).

These findings independently demonstrated how misregulation of neural progenitor cell proliferation can directly alter the brain size and consequently disrupt cognitive function. Notably, HMG and other megalencephaly/macrocephaly disorders are often associated with secondary anomalies in neuronal migration, whereas the primary microcephaly disorders often show no obvious brain anomalies other than simplification of the gyral pattern (Barkovich et al. 2012). This suggests that neuronal production and migration are highly synchronized processes during corticogenesis, and thereby the normal neuronal migration machineries cannot accommodate the misregulated excessive neuronal production. Such difference may partially explain why HMG patients exhibit much more prominent epileptic symptoms than MCPH patients (Flores-Sarnat 2002).

Genetic Mutational Mechanisms of Neurodevelopmental Disorders

From inherited to de novo mutations

Historically, conventional human genetics studies have primarily focused on inherited germline mutations as the basis of Mendelian diseases. Many rare and severe brain malformations are caused by autosomal recessive mutations carried by both healthy parents in the heterozygous state. Therefore, linkage analysis on the family pedigree has led to tremendous success in the identification genetic causes underlying these severe developmental disorders, with discoveries of essential genes involved in critical aspects of cortical development (Walsh & Engle 2010). In contrast, for more prevalent and milder conditions such as ASD, the genetic causes of the majority of cases (75-80%) remain unidentified following traditional methods (Miles 2011). Recent technological advances in high-resolution genomic microarrays and next-generation sequencing (NGS)

allow the identification of *de novo* or spontaneous germline mutations as important contributors of complex and sporadic neurodevelopmental disorders (Sanders et al. 2011; Levy et al. 2011; Sanders et al. 2012; State & Sestan 2012; Kong et al. 2012). Even more recently, the identification of several "brain-specific" somatic mosaic mutations as the genetic cause of brain malformations has further extended the spectrum of the mutational mechanisms of neurodevelopmental disorders (Poduri et al. 2012; Rivière et al. 2012; Lee et al. 2012).

De novo CNVs and SNVs

De novo mutations are defined as "new" genetic variants that are not present in either of the parents. They can arise in the germline of either parents, or post-zygotically during embryogenesis (Veltman & Brunner 2012; Vadlamudi et al. 2010). It has been widely speculated that rare sporadic genetic diseases are often caused by extremely deleterious *de novo* mutations, which would otherwise be under strong negative selective pressure due to the reproductive disadvantage of the patients (Eyre-Walker & Keightley 2007). More recently, a pioneering study carried out by Sebat, et al. (2007) show that the occurrence of *de novo* copy number variations (CNVs) is considerably elevated in patients with sporadic autism compared to healthy controls, suggesting *de novo* germline mutations as significant genetic contributors of complex diseases, particularly in neurodevelopmental disorders (Sebat et al. 2007).

With the widespread application of high-resolution genomic microarrays and next-generation sequencing in the recent years, a number of subsequent studies confirmed and refined the role of *de novo* mutations, including *de novo* CNVs and *de novo* single

nucleotide variants (SNVs), as important genetic risk factors of sporadic neurodevelopmental disorders (Vissers et al. 2010; O'Roak et al. 2011; Levy et al. 2011; Sanders et al. 2011; Gilman et al. 2011; Girard et al. 2011; Xu et al. 2011; Sanders et al. 2012; O'Roak et al. 2012; Neale et al. 2012; Iossifov et al. 2012; Mefford et al. 2010). Based on these large-scale genome-wide studies, the average *de novo* mutation rate for SNVs was estimated to be 1.2×10^{-8} per nucleotide per generation, resulting in 74 *de novo* germline SNVs in each individual's genome (Kong et al. 2012; Conrad et al. 2011). Additionally, *de novo* SNVs are found be predominately paternal in origin and age-dependent, suggesting replication errors as the major contributor to these mutations (Kong et al. 2012). *De novo* small indels and large CNVs (>100kb) are thought to occur at much lower frequency, estimated to be 3 per genome per generation and 1 out of 50 genomes per generation, respectively (Itsara et al. 2010; Fu et al. 2010). Despite their low frequency, large *de novo* CNVs collectively alter a larger fraction of the genome and are usually more deleterious and relevant to disease.

Neurodevelopmental disorders, including ID, ASD, SCZ and epilepsy, often manifest overlapping phenotypes, presumably as a consequence of general impairment of nervous system development. Remarkably, these disorders also share a highly similar large CNV landscape, with a number of recurrent CNVs being observed across diverse neurological phenotypes (Xu et al. 2008; Sharp et al. 2008; Mefford et al. 2008; de Vries et al. 2005; Sanders et al. 2011). These observations lead to a hypothesis that large CNVs tend to disrupt the homeostasis of normal neuronal development, which accounts for the increased CNV burden in a wide spectrum of neurodevelopmental and neuropsychiatric conditions (Coe et al. 2012). Rare and recurrent CNVs can be both inherited and *de novo*,

with CNV size positively correlated to the probability of arising *de novo* (Itsara et al. 2010). The molecular mechanisms underlying the formation of CNVs and structural variants (SVs) are currently under vigorous investigations and remain controversial. Three major mechanisms have been shown to account for the majority of CNVs: non-homologous end joining (NHEJ); replication-based mechanisms including fork stalling and template switching (FoSTeS) and microhomology mediated break-induced replication (MMBIR); and non-allelic homologous recombination (NAHR) (Zhang et al. 2009; Conrad et al. 2010; Mills et al. 2011; Kidd et al. 2010; Chiang et al. 2012). Each mechanism generates breakpoints with distinctive features and is associated with characteristic mutational "hotspots" that lead to the formation of recurrent CNVs (Fu et al. 2010; Itsara et al. 2009; Mills et al. 2011). Several large recurrent CNVs associated with neurological phenotypes are thought to occur at meiotic recombination hotspots with segmental duplications (SD) by NAHR (Turner et al. 2008; Itsara et al. 2012; Zhang et al. 2009; Veltman & Brunner 2012); and therefore, they primarily arise *de novo* during germline development. However, in rare cases, recurrent *de novo* CNVs can also happen at post-zygotic stage, resulting in special conditions named germline and somatic mosaicism (Koolen et al. 2012).

From germline to somatic mutations

De novo (meaning "new") mutations are operationally defined by as "new" genetic variants that are not detected in either of the parents. In fact, there are three different types of new mutations that fit under the umbrella term "*de novo*." Type I mutations arise in the germline of the parents due to either meiotic or mitotic errors to

give a single or a small number of mutated gametes. This type of mutation is transmitted to the child through a parent but the siblings are unlikely to be affected because the mutation would only present in a very limited number of gametes from the affected parent. Type II mutation arises at the post-zygotic stage during early embryonic development of the affected child, due to mitotic errors. This type of mutations is not transmitted from the parents and thereby is not detectable from the parents and is not recurrent among siblings. Depending upon when a type II mutation arises, it may affect only a fraction of cells from the whole body, a condition named as "*somatic mosaicism*" (**Figure 1-4**). Type III mutation arises during the post-zygotic development of the affected parent, resulting in a portion of germline cells carrying the mutation, and the mutation being transmitted to the child through an affected gamete. This condition is named as "*germline mosaicism*." Although the parent carries the mutation at a mosaic state, they are often misdiagnosed as non-affected due to the limitation of standard genetic testing on minor alleles that present at low level. Type III mutations are the only type of *de novo* mutations that lead to occasional recurrence among the siblings. Current studies on "*de novo*" mutations have primarily focused on type I mutations by identifying mutations not shared with the parents from simplex (non-recurrent) families (Sanders et al. 2012; Sanders et al. 2011; Iossifov et al. 2012; O'Roak et al. 2011; Xu et al. 2011; Levy et al. 2011). Type II and III mutations can also be occasionally detected and misclassified as type I mutations in these studies. The major

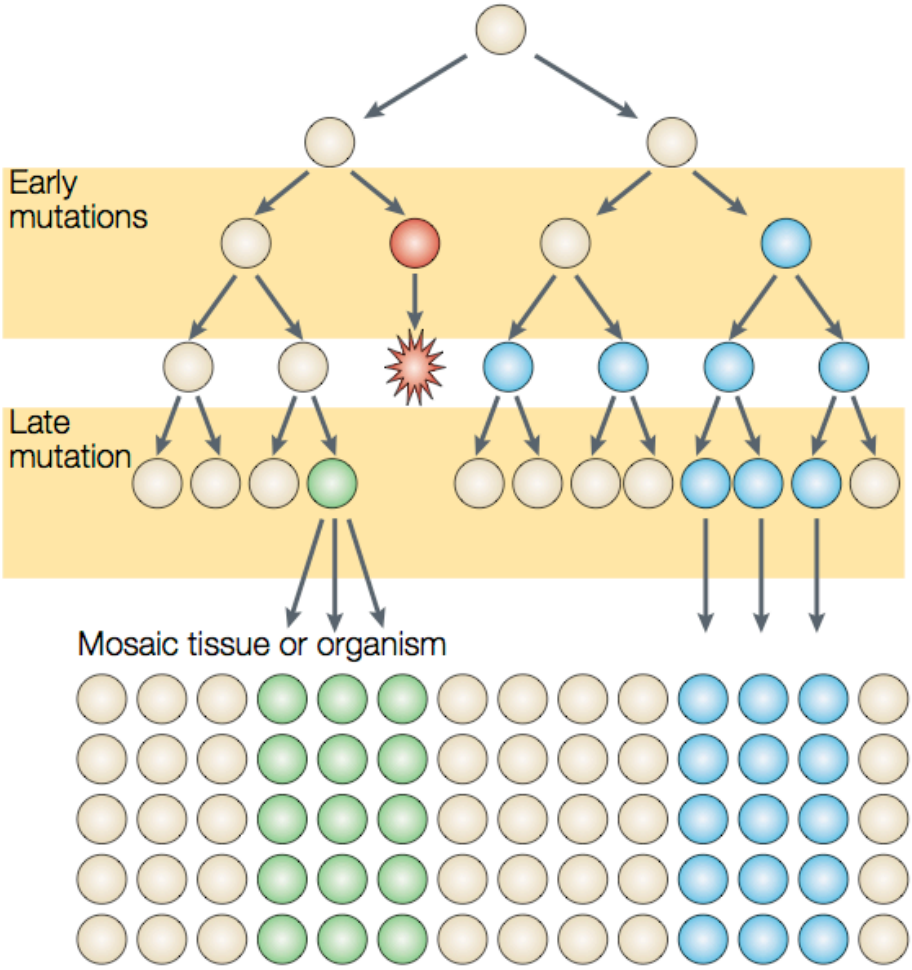


Figure 1-4. Schematic for the generation somatic mosaicism. Adapted from Youssoufian et al. (2002).
The percent of cells affected by a somatic mutation depends on its functional consequence on fitness as well as the time point when the mutation occurs.

focus of this dissertation is to develop effective methods to detect type II somatic mutations and to explore their contribution in the development of normal and diseased human brains.

Type II somatic mutations can be further classified into mutations arising during embryonic development and mutations arising postnatally throughout adulthood. Although with some overlap, the mutational mechanisms and functional consequences of somatic mutations arising at these two distinct developmental stages can be drastically different.

During embryogenesis, the single fertilized zygote is programmed to generate ~10 trillion cells that compose the human body through trillions of cell divisions (Erickson 2010). Genetic errors are inevitably introduced with each cell division due to the imperfections of DNA polymerases and the DNA damage repair machinery; therefore, we would expect our body to carry a tremendous number of somatic mutations and every cell in the body is likely to be different (Frumkin et al. 2005). Theoretical calculation predicts that every gene in the genome mutates many times throughout any individual's lifespan based on an estimated somatic mutation rate on the order of magnitude 10^{-7} -10^{-6} (Frank 2010). Despite the large number of somatic mutations present in our body, their functional consequences remain mostly uncharacterized. Similar to germline mutations, somatic mutations can be functionally neural, deleterious or advantageous. It is also important to note that for any given somatic mutation to have a significant functional effect at the organismal level, such mutation needs to be shared by a sufficient number of cells. Therefore, a somatic mutation with functional significance would either occur in a progenitor cell at very early stage of embryogenesis or result in a proliferative advantage

to the founder cell leading to its preferential clonal expansion (Erickson 2010). Additionally, deleterious mutations arising at early stages of embryogenesis are likely to be eliminated or diluted out during development. Several recent studies of chromosomal mosaicism in human preimplantation embryos highlight the prevalence of these early mutational events, suggesting that frequent somatic chromosomal instability may largely contribute to spontaneous miscarriages and constitutional chromosomal disorders, which were previously thought to originate primarily from the parental germline as type I *de novo* mutations (Vanneste et al. 2009; van Echten-Arends et al. 2011). On the other hand, despite the high prevalence of chromosomal mosaicism at the blastocyst stage, the noticeable preferential development of a diploid fetus from a mosaic embryo demonstrates that aneuploid cells at early embryonic stages are under strong negative selection, consistent with the predicted deleterious effect of aneuploidy (Eggan et al. 2002; van Echten-Arends et al. 2011). Taken together, the functional consequence of any somatic variant depends on both the time point at which it arises and its functional effect on the founder cell and its immediate progeny. The phenotypic manifestations of somatic mutations range from benign (e.g. birthmarks) to some most aggressive forms of human diseases (e.g. overgrowth syndromes and cancers).

Somatic mutations continue to arise postnatally, throughout individuals' lifespan. Similar to somatic mutations generated during embryo development, most postnatal somatic mutations arise through mitotic errors during tissue turnovers. Therefore, the abundance of somatic mutations generated postnatally highly depends on the tissue types and their turnover rates. For instance, skin epithelium and hematopoietic cells undergo rapid self-renewal, whereas the neurons that constitute 30% of the brain are strictly post-

mitotic with the exception of the hippocampus and olfactory system (Pellettieri & Alvarado 2007). Therefore, a much higher rate of replication-driven somatic mutations is expected in skin and hematopoietic system compared to brain. Examples of somatic mutations in normal individuals include moles growing on the skin, while recent studies confirm that somatic CNVs present at different frequencies between different tissues and they tend to accumulate with aging (Jacobs et al. 2012; Laurie et al. 2012; O'Huallachain et al. 2012). Somatic mutations can also occur in post-mitotic and quiescent cells due to flawed DNA damage repair. Neurons are the most unique type of post-mitotic cells, surviving for decades throughout our lifespan; therefore, it is expected that neurons would potentially accumulate more post-mitotic somatic mutations than other cells in the body. The genome instability of post-mitotic neurons is highlighted by observations that the CAG repeats causing Huntington's Disease (HD) continue to expand in post-mitotic neurons with focal somatic expansion of the CAG repeats often leading to early-onset of HD (Gonitel et al. 2008). The prevalence of age-related somatic mutation in normal brain, and its contribution to age-related neurological conditions, are important but technically challenging to explore with current technologies.

Somatic Mutations and Diseases

Although not commonly recognized, somatic mutations may contribute to a substantial portion of sporadic genetic diseases that are enriched for type I *de novo* germline mutations. In fact, many of the recurrent *de novo* mutations associated with neurodevelopmental disorders also arise as somatic mutations. For instance, heterozygous mutations in *SCN1A*, gene encoding the α1-subunit of the type I neuronal voltage-gated

sodium channel, are the primary genetic cause of Dravet syndrome, a severe, sporadic epileptic syndrome of infancy that is also associated with ID and ASD. 95% of *SCN1A* mutations identified are regarded as *de novo* (Vadlamudi et al. 2010; Gennaro et al. 2006). However, study of monozygotic twins with Dravet syndrome demonstrated that the *de novo* mutations could occur both in the parental germline—leading to concordant twins—or at post-zygotic stages resulting in discordant twins (Vadlamudi et al. 2010). Moreover, it is predicted that many of the most severely deleterious mutations can only exist as mosaic somatic mutations affecting a subset of cells in a human body, since they would be incompatible with life if affecting the whole body (Happle 1987). Therefore, somatic mutations exhibit a wider mutational spectrum than germline mutations, and may explain many disorders with unidentified genetic causes. This prediction is well supported by the observation that a number of chromosomal aneuploidy disorders other than trisomy 13, 18, 21 could only be tolerated as somatic mosaics (Erickson 2010). More recently, the genetic cause of Proteus syndrome, characterized by local overgrowth of skin, connective tissue and brain, was identified as a mosaic activating mutation in the *AKT1* gene, which results in a significant proliferative advantage of affected cells (Lindhurst et al. 2011). It is predicted that such mutation would otherwise be lethal in a non-mosaic state. This inspiring finding directly stimulated a number of additional studies identifying somatic activation mutations targeting the PI3K-AKT-mTOR pathway as the genetic causes of similar overgrowth syndromes (Hussain et al. 2011; Poduri et al. 2012; Lee et al. 2012; Rivière et al. 2012). In particular, our group identified the first "brain-only" somatic activating mutation in *AKT3* as one of the genetic causes of a brain overgrowth syndrome, hemimegalencephaly (HMG) (see previous section "Hemimegalencephaly"). Focal

lesions are a common characteristic feature of these somatic overgrowth syndromes; therefore, we hypothesize that additional brain malformation conditions involving focal lesions with unidentified genetic causes, such as focal cortical dysplasia (FCD)—an epileptic developmental brain lesion with neuropathological abnormalities suggestive of an early disruption of the normal brain development—are also likely to be ultimately explained by somatic mutations.

To further extend the importance of somatic mutation in the context of neurodevelopment, somatic mutations that do not confer proliferative advantages have also been previously identified, leading to local perturbations of normal brain development. For example, dominant mutation of *LIS1* and X-linked mutation in *DCX* in males can occur in a mosaic state to perturb the neuronal migration of a subset of neurons, resulting in the "double cortex" pattern, a milder form of lissencephaly (Sicca et al. 2003; Gleeson et al. 2000). Both the *LIS1* and *DCX1* somatic mutations were detectable from tissues outside of the brain, suggesting that the mutation occurred early in embryogenesis, before the neural lineage was specified. More importantly, the phenotypic manifestation of the disease depends on the percent mosaicism, consistent with the previous prediction that the earlier a somatic mutation arises during development, the more likely it is to cause a pathogenic condition (Gleeson et al. 2000; Erickson 2010). Furthermore, somatic mutations disrupting neurodevelopment or neuronal functions do not necessarily result in anatomically visible lesions. They could potentially also account for the complex idiopathic forms of neurological conditions, including non-syndromic ID and ASD, non-lesional epilepsy, and SCZ. As previously mentioned, *de novo* mutations in genes encoding ion channel subunits, such as *SCN1A* and *SCN2A*, are recurrently identified in

both syndromic and idiopathic (non-syndromic) forms of epilepsy and ASD, arising either at post-zygotic stage or in parental germline (O'Roak et al. 2011; Gennaro et al. 2006; O'Roak et al. 2012; Sanders et al. 2012). Similar to point mutations, recurrent CNVs that are known to arise *de novo* also occur both at post-zygotic stages and in the parental germline (Koolen et al. 2012). It is expected that mutations altering neuronal-specific functions, such as ion channels and neurotransmitter receptors, can severely impair cognitive functions while gross brain structure may remain intact. Mutations regulating neural-specific functions would not be under negative selection pressure during embryogenesis and would therefore be expected to be preserved if they occur early in development. However, the contribution of somatic mutations to sporadic neurodevelopmental and neuropsychiatric conditions is as of yet largely unexplored and should be a high priority of the field. It would also be interesting to determine whether the percent mosaicism of a somatic mutation correlates with the severity of the patient phenotype and the penetrance of a particular mutation. These studies will definitely add insight into the genetic heterogeneity as well as the variable penetrance and expressivity, of recurrent mutations identified in complex neurodevelopmental disorders. The major technical hurdle of addressing these question remains the limitation of current sequencing methods in detecting low-level mosaic somatic mutations.

The major challenge of detecting somatic variants from bulk samples (i.e. DNA extracted from a pool of affected and non-affected cells) is that the signal of the somatic variants is severely diluted out from the normal alleles present in the non-affected cells in the same sample. This problem becomes more pronounced when the percent mosaicism of the somatic variants gets low in the tissue being sampled. Current mutation detection

models assume that a mutant allele accounts for at least 50% of the signal, whereas somatic variants would only account for a much smaller fraction and would be likely regarded as signal noise that does not trigger a variant call. Tissue accessibility is the other major challenge for the detection of somatic mutations since the affected tissue, which presumably is most enriched for the mutation, may not always be available for genetic testing. For most neurological disorders, DNA specimens derived from blood are studied and brain tissues are largely unavailable; therefore it is predicted that most somatic mutations present in the brain but absent (or present at low-level) in blood would be missed. The development and implementation of non-invasive tissue sampling would be essential to address this issue.

Given the challenges, specific experimental designs are often required for studies of somatic mutations. For instance, studies that systematically survey a variety of tissue types led to the identification of somatic variants in both normal and diseased individuals (O'Huallachain et al. 2012; Kurek et al. 2012).With growing interest on studying low-level mosaic CNVs, several groups have recently developed bioinformatic tools to detect mosaic large CNVs from SNP/CNV hybrid microarrays by combining the subtle logRR change (log2 ratio measuring the copy number) with allelic imbalance indicated by the deviation of BAF score (B-Allele-Fraction) from expected heterozygous state (Jacobs et al. 2012; Laurie et al. 2012; Vattathil & Scheet 2013). These studies have shown that somatic mosaic CNVs exist in the blood of normal individuals and expand clonally. More interestingly, the frequency of detectable clonal mosaicism increases with age and is positively correlated with the risk of developing hematopoietic cancers (Jacobs et al. 2012; Laurie et al. 2012). Although these studies were able to detect large somatic CNVs

down to <10% percent mosaicism from bulk DNA, they are limited to a single class of mutation and are limited to the SNP chip platforms which are slowly getting replaced by next-generation sequencing (NGS) platforms. Compared to SNP chips, NGS allows the assessment of a full spectrum of mutations, including SNP, CNV/SV at higher resolution than SNP chips, retrotransposition insertions (L1, Alu and SVA) and microsatellites (MS) mutations; however, the tools available for studying somatic mutations from NGS data are still extremely lacking and have been barely applied to the detection of somatic mosaicism from normal tissues outside of cancers (Dewal et al. 2012). Furthermore, there will always be a technical detection limit of somatic mutations with low percent mosaicism from bulk samples when the signal level falls below the spontaneous sequencing error rate with the current sequencing technologies.

It is also important to note that somatic mutations arising postnatally are significantly more challenging to detect than those occurring at early embryonic stages because of their relative lack of clonality. Cells undergoing postnatal turnover are mostly terminally differentiated; therefore a somatic mutation acquired at a late stage will only affect a small number of cells with two exceptions: 1) if the mutation results in a proliferative advantage, which may lead to clonal expansion and often the development of benign tumors; 2) if a progenitor cell responsible for tissue turnover carries a somatic mutation either originating during embryonic development or acquired through flawed DNA damage repair during its quiescent stage, such a somatic mutation can be expanded clonally and become detectable by current technologies. Therefore, we choose to focus on somatic mutation arisen during embryonic development since they are technically easier to detect and highly relevant to neurodevelopmental disorders.

Somatic mosaic syndromes were first described by cytogeneticists decades ago using karyotyping to identify mosaic chromosomal disorders. The major advantage of karyotyping is its single cell resolution, which allows definitive separation and characterization of the small proportion of cells carrying the somatic mutations. However, this method is limited to the identification of large, microscopically visible chromosomal abnormalities. In order to systematically identify and quantify somatic mutations, whole genome sequencing of single cells would be necessary. Substantial progress has been made in single cell genomics in the past a few years, yielding valuable insight into the genetic heterogeneity of cancers and providing exciting future directions in characterizing somatic mosaicism of normal and disease human brains (Navin et al. 2011; Hou et al. 2012; Xu et al. 2011). These recent advances will be reviewed in the following section "*Overview of Recent Advances in Single-Cell Sequencing Technology and Applications*."

Somatic variants in normal brains

There are ~100 billions neurons in the human brain, comprising by different neuronal subtypes that are specified by morphology, gene expression, and patterns of connections. Furthermore, each neuron is estimated to form thousands of synapses with other neurons, and these synapses are responsible for each neuron's distinctive activities and functions, leading to the speculation that within each neuronal subtype, functional diversity further exists at the level of single neurons (Singer et al. 2010). Such an exquisite functional diversity has been proposed to account for our exceptional cognitive capacities. One of the major questions remains in neuroscience remains to understand the molecular basis of this functional diversity, which is together defined by intrinsic and

extrinsic factors. The proposed intrinsic attributes of neuronal diversity include the diversified expression patterns, epigenetic modifications, as well as genetic mosaicism (Muotri & Gage 2006). Although it is intriguing to propose that somatic variants positively contribute to the functional diversity of neurons, such a mechanism has to be balanced by the need for genome stability, as evident by the pathogenic somatic mutations causing neurological disorders (see previous section "*Somatic mutations and diseases*").

There have been long-standing hypotheses about the potential contribution of genetic variability in neurons to neuronal diversity but this topic remains highly controversial. In the 60's and 70's, several studies showed that Purkinje cells were tetraploid, which was further speculated to contribute to their distinct morphology and functions (Lapham 1968). However, these findings were later challenged and corrected in the late 70's and 80's as technical inconsistencies on DNA quantification (Mann et al. 1978; Swartz & Bhatnagar 1981). In the 90's, a series of studies suggested the potential involvement of the V(D)J recombination mechanism in the central nervous system (CNS) for genome diversification analogous to the immune system, stemming from the observation of brain expression of the recombination activating protein RAG-1, a key factor of the V(D)J recombination machinery (Chun et al. 1991). Furthermore, some human syndromes as well as mice mutants with DNA damage repair (DDR) defects only affect the immune and nervous system, suggesting some functional similarities between the two (Gao et al. 1998). However, studies of potential genetic recombination in brain were not thoroughly followed up or were proven otherwise using different experimental systems (Abeliovich et al. 1992). Extensive DNA rearrangements in neuronal genomes

were also hypothesized followed by the initial failure of cloning of viable mice from post-mitotic neurons (Hochedlinger & Jaenisch 2002). However, this was again disputed by the later success of cloning of a functionally normal mouse from an olfactory neuron (Eggan et al. 2004). Mechanisms of genetic diversity in the CNS continued to be explored and led to two recent hypotheses that somatic L1 retrotransposition and somatic aneuploidy may serve as somatic mutational mechanisms during neurogenesis to generate somatic variation and diversity (Singer et al. 2010; Rehen et al. 2005).

Human-specific LINE-1 (L1Hs) retrotransposons comprise the only known active autonomous transposon family in humans, with ~80-100 active L1Hs elements per individual (Hancks & Kazazian 2012). The activities of L1Hs have been mostly characterized in the germline, with a few recent reports of the occurrence of somatic L1Hs insertions in cancerous and normal cells (Miki et al. 1992; van den Hurk et al. 2007; Iskow et al. 2010; E. Lee et al. 2012a). However, a systematic characterization of L1Hs activity in somatic tissues is still lacking. Recent studies observed rare retrotransposition events of a L1Hs reporter gene in rodent brain *in vivo*, suggesting that L1Hs are active somatically during neurogenesis at some level in rodents (Muotri et al. 2005; Muotri et al. 2010). Furthermore, *in vitro* study of human neural progenitors estimated the rate of somatic L1Hs insertions to be around 10^{-5} events/cell based on the same L1Hs GFP-reporter used the *in vivo* studies on rodents (Coufal et al. 2009). However, an independent qPCR-based method on bulk DNA extracted from primary human brain tissues indirectly estimated a much higher rate of retrotransposon insertion of around 80-100 events/cell in hippocampus (Coufal et al. 2009). One additional study found evidence for widespread somatic L1Hs insertions in the human brain by targeted capture and deep sequencing

from bulk DNA (Baillie et al. 2011). However, these authors were unable to estimate a insertion rate due to the method limitation. Although intriguing, these studies are not entirely consistent with each other and are not direct proof of somatic retrotransposition activity since they failed to identify bona fide somatic insertions with their biological hallmark, namely target site duplication (TSD). Furthermore, the large discrepancy between the estimation of the retrotransposition rates *in vitro* and *in vivo* urges the use of an independent method that allows direct estimation of the insertion rate from human brain. Single cell analysis would achieve both goals by systematically identifying and quantifying somatic insertions from neurons directly isolated from human brains.

Aneuploidy defines a condition where the chromosome number within a cell deviates from the wildtype by part of a chromosome set. For instance, the human euploid genome contains 46 chromosomes, and the most common aneuploid condition in human is trisomy 21, the genetic cause of Down syndrome, which contains 47 chromosomes with an extra chromosome 21. Aneuploidy arises from chromosome mis-segregation during either mitosis or meiosis due to various mechanisms, including spindle-assembly checkpoint (SAC) defects, kinetochore and microtubule defects, and multipolar spindles (Siegel & Amon 2012). The functional consequence of aneuploid cells often includes slower proliferation, energy and proteotoxic stress, and genomic instability (Siegel & Amon 2012). Although generally detrimental, aneuploidy can also serve as a fast adaptive mechanism to improve fitness under selective pressure (Yona et al. 2012; Siegel & Amon 2012). Naturally occurring somatic aneuploidies have been reported from liver and brain, both resulted by multipolar mitoses (Rehen et al. 2001; Rehen et al. 2005; Duncan et al. 2010). Within human liver hepatocytes, greater than 25% were shown to be

aneuploid based on karyotyping; and the prevalence of aneuploid cells appears to associate with stress and increases with aging in mouse models (Duncan, Hanlon Newell, Smith, et al. 2012b; Duncan, Hanlon Newell, Bi, et al. 2012a). An increased rate of aneuploid cells has also been claimed in mice and human brain. It was initially observed in mouse neural progenitors of which ~33% were showed to be aneuploid, by spectral karyotyping (SKY). This rate substantially decreased in mature neurons, assayed by interphase FISH targeting the sex chromosomes, suggesting that the observed aneuploidy of progenitors confer proliferative disadvantages, premature differentiation or cell death (Rehen et al. 2001). Follow-up studies were carried out by the same group showing that the aneuploid progenitors arise from multipolar mitoses and that the aneuploid neurons observed in mice were functionally integrated into the neuronal circuitry (Yang et al. 2003; Kingsbury et al. 2005). A similar study was carried out to quantify the prevalence of aneuploidy in human brains; both neurons and glial cells from human brain tissue showed a slight increase in aneuploid frequency (5%) compared to lymphocytes control (0.6%) based on interphase FISH against chromosome 21 (Rehen et al. 2005). The authors further argued that this might be a significant underestimation of aneuploid cell frequencies since they only assayed for a single chromosome. These studies taken together generated a hypothesis that elevated levels of somatic neuronal aneuploidy may serve as a mechanism to generate genetic diversity within the human brain, analogous to the increased level of somatic retrotransposition (Muotri & Gage 2006).

Studies of neuronal aneuploidy in brain are subject to several technical concerns. First of all, interphase FISH is known to be prone to false positive results since the probes being used are usually targeting the alpha-satellite regions of different chromosomes and

they can falsely hybridize to other chromosomes due to rare variants shared by other chromosomes (Collin et al. 2009; Winsor et al. 1999). In fact, several follow-up studies from different groups failed to generate consistent results on the frequency of aneuploid neurons in human brain using interphase FISH targeting different genomic regions (Westra et al. 2008; Yurov et al. 2007; Thomas & Fenech 2008). An independent quantification method would be necessary to confirm the initial observations. Secondly, only a small subset of the genome (one or a few chromosomes) was assayed in the previous studies; therefore, the studies were unable to provide an accurate genome-wide representation of the "aneuploid" cells. For example, does an "aneuploid" neuron identified in these studies only have copy number gain or loss of the assayed chromosome, or is the neuron grossly aneuploid with gains or losses on multiple chromosomes? Single-cell copy number profiling from human brains provides a mean to address all these questions definitively.

All these studies on somatic variants in normal human brains are highly interesting and potentially high impact if they are proven to contribute to the functional diversification of neurons. However, several steps need to be taken to test the hypothesis. First of all, somatic variants are expected in every single somatic cell including neuron based on the previous discussion (see section "*From germline to somatic mutations*"). Therefore, identification of a certain type of somatic variant is not succifient to support the hypothesis that somatic variants are active generators of neuronal genetic diversity. A demonstration of significantly elevated levels of somatic variants compared to other tissues by more direct and quantitative means would be required to support the hypothesis. To date, V(D)J recombination in the immune system is the only known

mechanism to actively generate somatic genetic diversity through DNA rearrangement in lymphocytes, allowing a rapid adaptive response to new antigens (Jung et al. 2006). Such a mechanism is required for the proper function the of the immune system as evident by the immune-deficiency associated with V(D)J recombination defects (Jung et al. 2006). Therefore, if an analogous mechanism exists in the central nervous system, the disruption of it would be expected to lead to adverse effects on neuronal function. However, the identification of mammalian species that appear to have lost all L1 activity suggests that L1 retrotransposition may not be a universal requirement for mammalian neuronal function (Cantrell et al. 2008). Secondly, both retrotransposition and aneuploidy are regarded as stochastic mutational forces that generate random and mostly deleterious mutations as opposed to targeted V(D)J recombination. Both somatic retrotransposition and somatic aneuploidy are prevalent in cancer cells, highlighting the adverse effects of these two mutational forces on genome stability (E. Lee et al. 2012a; Weaver et al. 2007). Furthermore, the frequency of aneuploid neurons appears to increase in neurodegenerative diseases such as Alzhermier's disease (AD) and Ataxia-Telangiectasia (AT), further suggesting that aneuploidy may lead to cognitive dysfunction as opposed to beneficial functional diversity (Iourov et al. 2009).

Overview of Recent Advances in Single-Cell Sequencing Technology and Applications

Single-cell analysis is performed to understand cell-to-cell variability. Such variability has been systematically overlooked at the genomics level as most previous studies assumed an identical genome within the cells of an individual. With recent

technical advances, researchers have started to appreciate cell-to-cell differences at the genomic level and the biological significance of such genetic heterogeneity in the development of human diseases such as cancers (Frumkin et al. 2008; Navin et al. 2011; Xu et al. 2011; Hou et al. 2012). Significant progress has also been made on studies of embryonic development using a single-cell genomics approach to understand the genetic variability of early stage embryos (Vanneste et al. 2009; Yin et al. 2013). Studies on the hemizygous genome of human single sperms have reveal remarkable insight into germline development regarding the homologous recombination patterns as well as *de novo* point mutation rates (Wang et al. 2012; Lu et al. 2012). Moreover, single-cell genomics has benefitted tremendously microbial biology by identifying uncultivable new microbial species as a mean to study ecosystem evolution (Blainey & Quake 2011; Swan et al. 2011). In addition to single-cell genomics, remarkable progress has also been made in gene expression analyses and transcriptome studies at the single-cell level, revealing complex biological regulatory networks and identifying new cell types (Ramsköld et al. 2012; Hashimshony et al. 2012). Despite all this recent progress, all current single cell studies are limited because they rely on a secondary amplification step because of the difficulty of directly analyzing the extremely tiny amount (6pg) of genetic material from a single cell. The secondary amplification inevitably introduces biases and artifacts, setting limitations on the applications of these methods.

The two major challenges to single-cell genomic analysis are: 1) the isolation of single cells with specific phenotypes of interest; and 2) whole-genome or whole transcriptome amplification from the isolated single cells. Current means for single cell isolation include fluorescence-activated cell sorting (FACS), laser-capture

microdissection (LCM) and microfluidics. FACS has been the most reliable method for single cell isolation, featuring the highest throughput and allowing for the use of various immunofluorescent markers to isolate specific cell populations. However, FACS relies heavily on the availability and quality of antibodies and requires a relatively large number of cells to start with, making it difficult to study rare and/or unknown cell populations (Navin & Hicks 2011; Kalisky et al. 2011). LCM is robust in permitting the identification of specific cell populations based on both morphology and a wider spectrum of antibodies (Tietjen et al. 2003). However, LCM is low-throughput, and therefore not easily applicable to studies of large numbers of cells. Microfluidic systems have been recently developed and incorporated into single-cell -omics studies and promise a few major advantages. First of all, as closed systems, microfluidic systems effectively minimizes the risk of external contamination, and allows for the isolation of cells and amplification of genetic material to take place in the same chamber, greatly simplifying the work flow (Tietjen et al. 2003; Zare & Kim 2010). Secondly, microfluidic systems allow an extremely small volumes for secondary amplification, which significantly brings down the cost and helps to reduce amplification bias (Marcy et al. 2007). The major limitation of exsiting microfluidic systems is their accessibility, as it requires certain expertise for its operation and maintenance. However, with commercially available microfluidic systems now being installed at a growing number of research institutes, this drawback is expected to be overcome in a near future.

A second important challenge of single-cell -omics studies is amplification bias. All currently used amplification methods introduce amplification artifacts such as errors, chimeras and false-positive copy number changes to varying degrees (Wang et al. 2012;

Blainey & Quake 2011; Lasken & Stockwell 2007); therefore, careful characterization of each method and its limitations is essential for critical analysis and interpretation of any single-cell -omics datasets. Depending upon the applications, different amplification methods may be applied (Bergen et al. 2005); therefore, careful evaluation and comparison of currently available amplification methods are important for achieving the best results. Eventually, more advanced single cell amplification methods that amplify the genome linearly with higher fidelity will need to be developed.

With the expansion of next generation sequencing technologies, an increasing number of clinical applications have also been proposed for the use of single-cell sequencing. Preimplantation genetic diagnosis (PGD) for IVF has pioneered the use of single-cell genomics to screen for cell aneuploidy, first using array-based methods and now moving to next-generation sequencing based (Vanneste et al. 2009; Yin et al. 2013). In the near future, it is expected that single-cell sequencing technology would be widely applied to noninvasive prenatal diagnosis through analyzing circulating fetal cells from maternal plasma. Additionally, single cell haplotyping techniques are being actively developed, providing tremendous power for the understanding of germline meiotic recombinations (Fan et al. 2011). Similarly, single-cell sequencing can be applied to the identification and analysis of circulating cancer cells (CTC), providing valuable information on cancer diagnosis, prognosis as well as treatments (Navin & Hicks 2011). The major technical challenges of these clinical applications remains the cost and consistency of current methods. With the further reduction of sequencing cost in the near future and the maturation of single-cell technology, we would expect rapid advances in

medical applications allowing for more sensitive diagnosis using the single-cell technology.

References

Abeliovich, A. et al., 1992. On somatic recombination in the central nervous system of transgenic mice. *Science (New York, N.Y.)*, 257(5068), pp.404–410.

Al-Dosari, M.S. et al., 2010. Novel CENPJ mutation causes Seckel syndrome. *Journal of medical genetics*, 47(6), pp.411–414.

Baillie, J.K. et al., 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, 479(7374), pp.534–537.

Barkovich, A.J. et al., A developmental and genetic classification for malformations of cortical development: update 2012.

Bergen, A.W. et al., 2005. Comparison of yield and genotyping performance of multiple displacement amplification and OmniPlex™ whole genome amplified DNA generated from multiple DNA sources. *Human mutation*, 26(3), pp.262–270.

Bettencourt-Dias, M. et al., 2011. Centrosomes and cilia in human disease. *Cell*, 27(8), pp.307–315.

Bilguvar, K. et al., 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, 467(7312), pp.207–210.

Blainey, P.C. & Quake, S.R., 2011. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic acids research*, 39(4), p.e19.

Breunig, J.J., Haydar, T.F. & Rakic, P., 2011. Neural Stem Cells: Historical Perspective and Future Prospects. *Neuron*, 70(4), pp.614–625.

Cantrell, M.A. et al., 2008. Loss of LINE-1 activity in the megabats. *Genetics*, 178(1), pp.393–404.

Chiang, C. et al., 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature Genetics*, 44(4), pp.390–397.

Chun, J.J.M. et al., 1991. The recombination activating gene-1 (RAG-1) transcript is present in the murine central nervous system. *Cell*, 64(1), pp.189–200.

Coe, B.P., Girirajan, S. & Eichler, E.E., 2012. The genetic variability and commonality of neurodevelopmental disease C. E. Schwartz & G. Neri, eds. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 160C(2), pp.118–129.

Collin, A., Sladkevicius, P. & Soller, M., 2009. False-positive prenatal diagnosis of trisomy 18 by

interphase FISH: hybridization of chromosome 18 alpha-satellite probe (D18Z1) to chromosome 2. *Prenatal diagnosis*, 29(13), pp.1279–1281.

Conrad, D.F. et al., 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature Genetics*, 42(5), pp.385–391.

Conrad, D.F. et al., 2011. Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7), pp.712–714.

Coufal, N.G. et al., 2009. L1 retrotransposition in human neural progenitor cells. *Nature*, 460(7259), pp.1127–1131.

de Vries, B.B.A. et al., 2005. Diagnostic genome profiling in mental retardation. *American journal of human genetics*, 77(4), pp.606–616.

Dewal, N. et al., 2012. Calling amplified haplotypes in next generation tumor sequence data. *Genome research*, 22(2), pp.362–374.

Duncan, A.W. et al., 2010. The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature*, 467(7316), pp.707–710.

Duncan, A.W., Hanlon Newell, A.E., Bi, W., et al., 2012a. Aneuploidy as a mechanism for stress-induced liver adaptation. *The Journal of clinical investigation*, 122(9), pp.3307–3315.

Duncan, A.W., Hanlon Newell, A.E., Smith, L., et al., 2012b. Frequent aneuploidy among normal human hepatocytes. *Gastroenterology*, 142(1), pp.25–28.

Eggan, K. et al., 2002. Male and female mice derived from the same embryonic stem cell clone by tetraploid embryo complementation. *Nature biotechnology*, 20(5), pp.455–459.

Eggan, K. et al., 2004. Mice cloned from olfactory sensory neurons. *Nature*, 428(6978), pp.44–49.

Engelman, J.A., Luo, J. & Cantley, L.C., 2006. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nature Reviews Genetics*, 7(8), pp.606–619.

Erickson, R.P., 2010. Somatic gene mutation and human disease other than cancer: An update. *Mutation Research/Reviews in Mutation Research*, 705(2), pp.96–106.

Eyre-Walker, A. & Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8), pp.610–618.

Fan, H.C. et al., 2011. Whole-genome molecular haplotyping of single cells. *Nature biotechnology*, 29(1), pp.51–57.

Flores-Sarnat, L., 2002. Hemimegalencephaly: part 1. Genetic, clinical, and imaging aspects. *Journal of child neurology*, 17(5), pp.373–84– discussion 384.

Flores-Sarnat, L. et al., 2003. Hemimegalencephaly: part 2. Neuropathology suggests a disorder of cellular lineage. *Journal of child neurology*, 18(11), pp.776–785.

Frank, S.A., 2010. Colloquium Paper: Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences*, 107(suppl_1), pp.1725–1730.

Frumkin, D. et al., 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS computational biology*, 1(5), p.e50.

Frumkin, D. et al., 2008. Cell Lineage Analysis of a Mouse Tumor. *Cancer research*, 68(14), pp.5924–5931.

Fu, W. et al., 2010. Identification of copy number variation hotspots in human populations. *American journal of human genetics*, 87(4), pp.494–504.

Gao, Y. et al., 1998. A Critical Role for DNA End-Joining Proteins in Both Lymphogenesis and Neurogenesis. *Cell*, 95(7), pp.891–902.

Gennaro, E. et al., 2006. Somatic and germline mosaicisms in severe myoclonic epilepsy of infancy. *Biochemical and biophysical research communications*, 341(2), pp.489–493.

Gilman, S.R. et al., 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70(5), pp.898–907.

Gilmore, E.C. & Walsh, C.A., 2012. Genetic causes of microcephaly and lessons for neuronal development. *Wiley Interdisciplinary Reviews: Developmental Biology*, pp.n/a–n/a.

Girard, S.L. et al., 2011. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics*, 43(9), pp.860–863.

Gleeson, J.G. et al., 1998. Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. *Cell*, 92(1), pp.63–72.

Gleeson, J.G. et al., 2000. Somatic and germline mosaic mutations in the doublecortin gene are associated with variable phenotypes. *American journal of human genetics*, 67(3), pp.574–581.

Gonitel, R. et al., 2008. DNA instability in postmitotic neurons. *Proceedings of the National Academy of Sciences*, 105(9), pp.3467–3472.

Griffith, E. et al., 2008. Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. *Nature Genetics*, 40(2), pp.232–236.

Hancks, D.C. & Kazazian, H.H., 2012. Active human retrotransposons: variation and disease. *Current opinion in genetics & development*, 22(3), pp.191–203.

Hansen, D.V. et al., 2010. Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature*, 464(7288), pp.554–561.

Hashimshony, T. et al., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports*, 2(3), pp.666–673.

Happle, R., 1987. Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects

involving the skin. *Journal of the American Academy of Dermatology*, 16(4), pp.899–906.

Hawrylycz, M.J. et al., 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416), pp.391–399.

Hochedlinger, K. & Jaenisch, R., 2002. Nuclear transplantation: lessons from frogs and mice. *Current Opinion in Cell Biology*, 14(6), pp.741–748.

Hou, Y. et al., 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 148(5), pp.873–885.

Hussain, K. et al., 2011. An activating mutation of AKT2 and human hypoglycemia. *Science (New York, N.Y.)*, 334(6055), p.474.

Iossifov, I. et al., 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2), pp.285–299.

Iourov, I.Y. et al., 2009. Aneuploidy in the normal, Alzheimer's disease and ataxia-telangiectasia brain: differential expression and pathological meaning. *Neurobiology of disease*, 34(2), pp.212–220.

Iskow, R.C. et al., 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7), pp.1253–1261.

Itsara, A. et al., 2010. De novo rates and selection of large copy number variation. *Genome research*, 20(11), pp.1469–1481.

Itsara, A. et al., 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *American journal of human genetics*, 84(2), pp.148–161.

Itsara, A. et al., 2012. Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *American journal of human genetics*, 90(4), pp.599–613.

Jacobs, K.B. et al., 2012. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics*, 44(6), pp.651–658.

Jung, D. et al., 2006. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology*, 24, pp.541–570.

Kalay, E. et al., 2011. CEP152 is a genome maintenance protein disrupted in Seckel syndrome. *Nature Genetics*, 43(1), pp.23–26.

Kalisky, T. & Quake, S.R., 2011. Single-cell genomics. *Nature methods*, 8(4), pp.311–314.

Kalisky, T., Blainey, P. & Quake, S.R., 2011. Genomic analysis at the single-cell level. *Annual Review of Neuroscience*, 45, pp.431–445.

Kidd, J.M. et al., 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5), pp.837–847.

Kiecker, C. & Lumsden, A., 2012. The Role of Organizers in Patterning the Nervous System. *Annual*

Review of Neuroscience, 35(1), pp.347–367.

Kingsbury, M.A. et al., 2005. Aneuploid neurons are functionally active and integrated into brain circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17), pp.6143–6147.

Kong, A. et al., 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412), pp.471–475.

Konopka, G. et al., 2012. Human-specific transcriptional networks in the brain. *Neuron*, 75(4), pp.601–617.

Konopka, G. et al., 2009. Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, 462(7270), pp.213–217.

Koolen, D.A. et al., 2012. Two families with sibling recurrence of the 17q21.31 microdeletion syndrome due to low-grade mosaicism. *European journal of human genetics : EJHG*, 20(7), pp.729–733.

Kurek, K.C. et al., 2012. Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. *American journal of human genetics*, 90(6), pp.1108–1115.

LaMonica, B.E. et al., 2012. OSVZ progenitors in the human cortex: an updated perspective on neurodevelopmental disease. *Trends in Neurosciences*, 22(5), pp.747–753.

Lapham, L.W., 1968. Tetraploid DNA content of Purkinje neurons of human cerebellar cortex. *Science (New York, N.Y.)*, 159(3812), pp.310–312.

Lasken, R.S. & Stockwell, T.B., 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology*, 7, p.19.

Laurie, C.C. et al., 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, 44(6), pp.642–650.

Lee, E. et al., 2012a. Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, 337(6097), pp.967–971.

Lee, J.H. et al., 2012b. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nature Genetics*, 44(8), pp.941–945.

Levy, D. et al., 2011. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5), pp.886–897.

Lindhurst, M.J. et al., 2011. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *The New England journal of medicine*, 365(7), pp.611–619.

Lu, S. et al., 2012. Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing. *Science (New York, N.Y.)*, 338(6114), pp.1627–1630.

Lui, J.H., Hansen, D.V. & Kriegstein, A.R., 2011. Development and evolution of the human neocortex. *Cell*, 146(1), pp.18–36.

Mann, D.M., Yates, P.O. & Barton, C.M., 1978. The DNA content of Purkinje cells in mammals. *The Journal of comparative neurology*, 180(2), pp.345–347.

Manzini, M.C. & Walsh, C.A., 2011. What disorders of cortical development tell us about the cortex: one plus one does not always make two. *Current opinion in genetics & development*, 21(3), pp.333–339.

Marcy, Y. et al., 2007. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS genetics*, 3(9), pp.1702–1708.

Marshall, C.R. et al., 2008. Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics*, 82(2), pp.477–488.

Mefford, H.C. et al., 2010. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS genetics*, 6(5), p.e1000962.

Mefford, H.C. et al., 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *The New England journal of medicine*, 359(16), pp.1685–1699.

Miki, Y. et al., 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research*, 52(3), pp.643–645.

Miles, J.H., 2011. Autism spectrum disorders--a genetics review. *Genetics in Medicine*, 13(4), pp.278–294.

Mills, R.E. et al., 2011. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332), pp.59–65.

Molyneaux, B.J. et al., 2007. Neuronal subtype specification in the cerebral cortex. *Nature Genetics*, 8(6), pp.427–437.

Muotri, A.R. & Gage, F.H., 2006. Generation of neuronal variability and complexity. *Nature*.

Muotri, A.R. et al., 2010. L1 retrotransposition in neurons is modulated by MeCP2. *Nature*.

Muotri, A.R. et al., 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, 435(7044), pp.903–910.

Navin, N. & Hicks, J., 2011. Future medical applications of single-cell sequencing in cancer. *Genome medicine*, 3(5), p.31.

Navin, N. et al., 2011. Tumour evolution inferred by single-cell sequencing. *Nature Genetics*, 472(7341), pp.90–94.

Neale, B.M. et al., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397), pp.242–245.

Nicholas, A.K. et al., 2010. WDR62 is associated with the spindle pole and is mutated in human microcephaly. *Nature Genetics*, 42(11), pp.1010–1014.

O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), pp.585–589.

O'Roak, B.J. et al., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397), pp.246–250.

O'Huallachain, M. et al., 2012. Extensive genetic variation in somatic human tissues. *Proceedings of the National Academy of Sciences*, 109(44), pp.18018–18023.

Pellettieri, J. & Alvarado, A.S., 2007. Cell Turnover and Adult Tissue Homeostasis: From Humans to Planarians. *Annual Review of Neuroscience*, 41(1), pp.83–105.

Poduri, A. et al., 2012. Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron*, 74(1), pp.41–48.

Rakic, P., 2009. Evolution of the neocortex: a perspective from developmental biology. *Nature Genetics*, 10(10), pp.724–735.

Ramsköld, D. et al., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, 30(8), pp.777–782.

Rehen, S.K. et al., 2001. Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), pp.13361–13366.

Rehen, S.K. et al., 2005. Constitutional aneuploidy in the normal human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(9), pp.2176–2180.

Reillo, I. et al., 2011. A role for intermediate radial glia in the tangential expansion of the mammalian cerebral cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 21(7), pp.1674–1694.

Rivière, J.-B. et al., 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nature Genetics*, 44(8), pp.934–940.

Sanders, S.J. et al., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397), pp.237–241.

Sanders, S.J. et al., 2011. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, 70(5), pp.863–885.

Swan, B.K. et al., 2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science (New York, N.Y.)*, 333(6047), pp.1296–1300.

Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823), pp.445–449.

Sharp, A.J. et al., 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, 40(3), pp.322–328.

Shen, J. et al., 2010. Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nature Genetics*, 42(3), pp.245–249.

Sicca, F. et al., 2003. Mosaic mutations of the LIS1 gene cause subcortical band heterotopia. *Neurology*,

61(8), pp.1042–1046.

Siegel, J.J. & Amon, A., 2012. New insights into the troubles of aneuploidy. *Annual review of cell and developmental biology*, 28, pp.189–214.

Singer, T. et al., 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in Neurosciences*, 33(8), pp.345–354.

State, M.W. & Sestan, N., 2012. The Emerging Biology of Autism Spectrum Disorders. *Science (New York, N.Y.)*, 337(6100), pp.1301–1303.

Sullivan, P.F., Daly, M.J. & O'Donovan, M., 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, 13(8), pp.537–551.

Swartz, F.J. & Bhatnagar, K.P., 1981. Are CNS neurons polyploid? A critical analysis based upon cytophotometric study of the DNA content of cerebellar and olfactory bulbar neurons of the bat. *Brain research*, 208(2), pp.267–281.

Thomas, P. & Fenech, M., 2008. Chromosome 17 and 21 aneuploidy in buccal cells is increased with ageing and in Alzheimer's disease. *Mutagenesis*, 23(1), pp.57–65.

Thornton, G.K. & Woods, C.G., 2009. Primary microcephaly: do all roads lead to Rome? *Cell*, 25(11), pp.501–510.

Tietjen, I. et al., 2003. Single-cell transcriptional analysis of neuronal progenitors. *Neuron*, 38(2), pp.161–175.

Turner, D.J. et al., 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, 40(1), pp.90–95.

Vadlamudi, L. et al., 2010. Timing of de novo mutagenesis--a twin study of sodium-channel mutations. *The New England journal of medicine*, 363(14), pp.1335–1340.

van den Hurk, J.A.J.M. et al., 2007. L1 retrotransposition can occur early in human embryonic development. *Human molecular genetics*, 16(13), pp.1587–1592.

van Echten-Arends, J. et al., 2011. Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Human reproduction update*, 17(5), pp.620–627.

Vanneste, E. et al., 2009. Chromosome instability is common in human cleavage-stage embryos. *Nature medicine*, 15(5), pp.577–583.

Vattathil, S. & Scheet, P., 2013. Haplotype-based profiling of subtle allelic imbalance with SNP arrays. *Genome research*, 23(1), pp.152–158.

Veltman, J.A. & Brunner, H.G., 2012. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8), pp.565–575.

Vissers, L.E.L.M. et al., 2010. A de novo paradigm for mental retardation. *Nature Genetics*, 42(12), pp.1109–1112.

Walsh, C.A. & Engle, E.C., 2010. Allelic diversity in human developmental neurogenetics: insights into biology and disease. *Neuron*, 68(2), pp.245–253.

Wang, J. et al., 2012. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell*, 150(2), pp.402–412.

Weaver, B.A.A. et al., 2007. Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell*, 11(1), pp.25–36.

Westra, J.W. et al., 2008. Aneuploid mosaicism in the developing and adult cerebellar cortex. *The Journal of comparative neurology*, 507(6), pp.1944–1951.

Winsor, E.J. et al., 1999. Risk of false-positive prenatal diagnosis using interphase FISH testing: hybridization of alpha-satellite X probe to chromosome 19. *Prenatal diagnosis*, 19(9), pp.832–836.

Wynshaw-Boris, A. et al., 2010. Lissencephaly: mechanistic insights from animal models and potential therapeutic strategies. *Seminars in cell & developmental biology*, 21(8), pp.823–830.

Xu, B. et al., 2011. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genetics*, 43(9), pp.864–868.

Xu, B. et al., 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics*, 40(7), pp.880–885.

Yang, A.H. et al., 2003. Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(32), pp.10454–10462.

Yin, X. et al., 2013. Massively Parallel Sequencing for Chromosomal Abnormality Testing in Trophectoderm Cells of Human Blastocysts. *Biology of reproduction*.

Yona, A.H. et al., 2012. Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences*, 109(51), pp.21010–21015.

Yu, T.W. et al., 2010. Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. *Nature Genetics*, 42(11), pp.1015–1020.

Yurov, Y.B. et al., 2007. Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PloS one*, 2(6), p.e558.

Zare, R.N. & Kim, S., 2010. Microfluidic platforms for single-cell analysis. *Annual review of biomedical engineering*.

Zecevic, N., Hu, F. & Jakovcevski, I., 2011. Interneurons in the developing human neocortex. *Developmental neurobiology*, 71(1), pp.18–33.

Zeng, H. et al., 2012. Large-Scale Cellular-Resolution Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures. *Cell*, 149(2), pp.483–496.

Zhang, F. et al., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10, pp.451–481.

Chapter 2: Inherited Recessive Germline Mutations in Brain Malformations

This chapter contains work from the manuscript "Human Mutations in *NDE1* Cause Extreme Microcephaly with Lissencephaly", published in *American Journal of Human Genetics*, May 13, 2011; 88:536-547. The text and figures were modified to fit the format of this dissertation. Xuyu Cai was co-first author of the manuscript with Fowzan Alkuraya. One of the two families studied (Family 1) was contributed by collaborator Fowzan Alkuraya. Clinical characterization was done by Ganesh Mochida, Brenda Barry and Jennifer Partlow. Homozygosity mapping and Sanger sequencing was done with help from Sean Hill and Jillian Felie. Functional characterization of both mutant alleles was solely done by Xuyu Cai. Immunocytochemistry of *Nde1*$^{-/-}$ MEFs was done by Carina Emery and Yuanyi Feng. Immunocytochemistry of patient lymphoblasts was done by Fowzan Alkuraya.

Summary

Genes disrupted in human microcephaly ("small brain") define key regulators of neural progenitor proliferation and cell fate specification. In comparison, genes mutated in human lissencephaly (lissos, "smooth," cephalos, "brain") highlight critical regulators of neuronal migration. Here we report two families with extreme microcephaly and grossly simplified cortical gyral structure, an entity referred to as microlissencephaly, and show that they carry homozygous frameshift mutations in *NDE1*. *NDE1* encodes a multidomain protein that localizes to the centrosome and mitotic spindle poles. Both human mutations in *NDE1* truncate the C-terminal domains of NDE1, which are essential for interactions with cytoplasmic dynein to regulate cytoskeletal dynamics in mitosis, and for phosphorylation of NDE1 by CDK1 in a cell cycle-dependent fashion. We show that the patient NDE1 proteins are unstable, cannot bind cytoplasmic dynein and do not localize properly to the centrosome. The role of NDE1 in cell cycle progression likely contributes to the profound neuronal proliferation defects evident in *Nde1* null mice and patients with *NDE1* mutations, demonstrating the essential role of *NDE1* in human cerebral cortical neurogenesis.

Introduction

The exquisitely organized formation of cerebral cortical neurons from the cortical neuroepithelium has provided an important system for studying the control of cell proliferation and cell fate. Cortical progenitors, forming a pseudostratified epithelium with nuclei in the ventricular germinal zone, have the capacity for symmetrical cell

divisions to form two dividing daughter cells, or asymmetrical cell divisions to generate one dividing daughter as well as post-mitotic neurons that populate the developing cortex. After exiting the cell cycle, post-mitotic neurons migrate away from the ventricular zone to the incipient cerebral cortical layers to establish the highly organized cortical architecture. Human autosomal recessive primary microcephaly (MCPH, [MIM251200]), or microcephaly vera, manifests with small but architecturally fairly normal brains. Multiple human genes mutated in MCPH encode proteins that localize to the centrosome and/or mitotic spindle poles (Thornton & Woods 2009). Many of them have been implicated in regulating progenitor cell cycle progression and the decision of progenitors to continue proliferating or to differentiate into post-mitotic neurons (Thornton & Woods 2009). In contrast to microcephaly, human lissencephaly manifests with a simplified cortical gyration pattern reflecting abnormal histological organization of the cortical layers, but normal brain volume, and most often reflects disruption of neuronal migration (Wynshaw-Boris et al. 2010). Identified genetic causes of lissencephaly include mutations in *LIS1* [MIM 601545], *DCX* [MIM 300121], *RELN* [MIM 600514] and *TUBA1A* [MIM 602519] (Dobyns et al. 1993; Gleeson et al. 1998; Hong et al. 2000; Poirier et al. 2007). Recently, mutations in *WDR62* [MIM613583] have been associated with microcephaly with a variety of architectural defects of the cortex as well (Nicholas et al. 2010; Yu et al. 2010; Bilguvar et al. 2010), suggesting additional overlap in the genes that regulate proliferation and migration. But there have long been rare cases of a reduction in brain size typical of microcephaly that is also associated with simplification of cerebral cortical gyration that falls within the spectrum of lissencephaly. These two findings together have been referred to as “microlissencephaly”, but the genetic and

mechanistic causes of microlissencephaly are unknown (Norman et al. 1976; Dobyns et al. 1984).

Nuclear distribution E (NudE) was originally identified in *Aspergillus nidulans* as an essential regulator of a common nuclear migration pathway, which also involves NudC (DYNC1H1, [MIM600112]) and NudF (LIS1) (Efimov & Morris 2000). The mammalian orthologues of NudE include *NDE1* [MIM 609449] and *NDE1-Like 1* (*NDEL1*, [MIM 607538]). Mouse *Nde1* is highly expressed in cortical neural progenitors and encodes a protein that localizes to the centrosome and mitotic spindle poles. It is also known to physically interact with the cytoplasmic dynein complex and Lis1 (Feng et al. 2000; Feng & Walsh 2004; McKenney et al. 2010). The dynein-Lis1-Nde1 complex has an essential role in cytoskeleton dynamics in a wide range of cellular processes, including mitosis, nuclei positioning, and cell migration (Wynshaw-Boris et al. 2010; Wynshaw-Boris 2007). Loss of *Nde1* in mouse models causes profound defects in cerebral corticogenesis but only modest defects in neuronal migration (Feng & Walsh 2004). *Nde1*-null cortical progenitors showed defects in centrosome duplication and mitotic spindle assembly, which resulted in severe mitotic arrest/delay, spindle mis-orientation and mis-positioning of the mitotic chromosomes. These mitotic defects were thought to account for the premature depletion of progenitor cells in *Nde1*-null brains through impaired cell cycle progression, and pre-mature cell cycle exit by progenitors to form neurons (Feng & Walsh 2004).

In this study, we demonstrate that human *NDE1* mutations cause a severe microlissencephaly syndrome, resembling that described by Norman and Roberts

(Norman et al. 1976; Dobyns et al. 1984). The identified frameshift mutations result in truncation of the C-terminal domains and disruption of several key functions of NDE1.

Results

Clinical Characterization of Microlissencephaly

Ongoing efforts to characterize the genetic bases of microcephaly by a large collaborative effort known as the Microcephaly Collaborative identified two families whose children share remarkably severe microcephaly (head circumferences more than -11 to -13 standard deviations below the mean for age), abnormal cortical gyration, short stature (more than -2 and -4 standard deviations below the mean for age), microsomia (weight more than -2 and -5 standard deviations below the mean for age) and a prominent broad nasal bridge (Figure 1; see Supplementary Clinical Information).

Family 1 originates from Eastern Saudi Arabia. The parents are healthy and are reported to be first cousins. They have two daughters affected with extreme microcephaly (**Figure 2-1a**). The first affected daughter (08DG00535; Family1, IV-1 in **Figure 2-1a**) was first evaluated at the age of 33 months, at which time her head circumference was 34.4cm, length was 81.4 cm and weight was 8.6 kg (8.9, 3 and 3.8 standard deviations below the mean respectively). Physical examination at that time was notable for marked hypertonia and global developmental delay. MRI imaging revealed severe microcephaly, with proportionate reduction in the size of most other brain structures, including the cerebellum and brain stem, associated with agenesis of the corpus callosum. The gyral folding of the cerebral cortex was extremely simplified, with almost no detectable sulci other than the Sylvian fissure. At the age of seven years, she displayed extreme

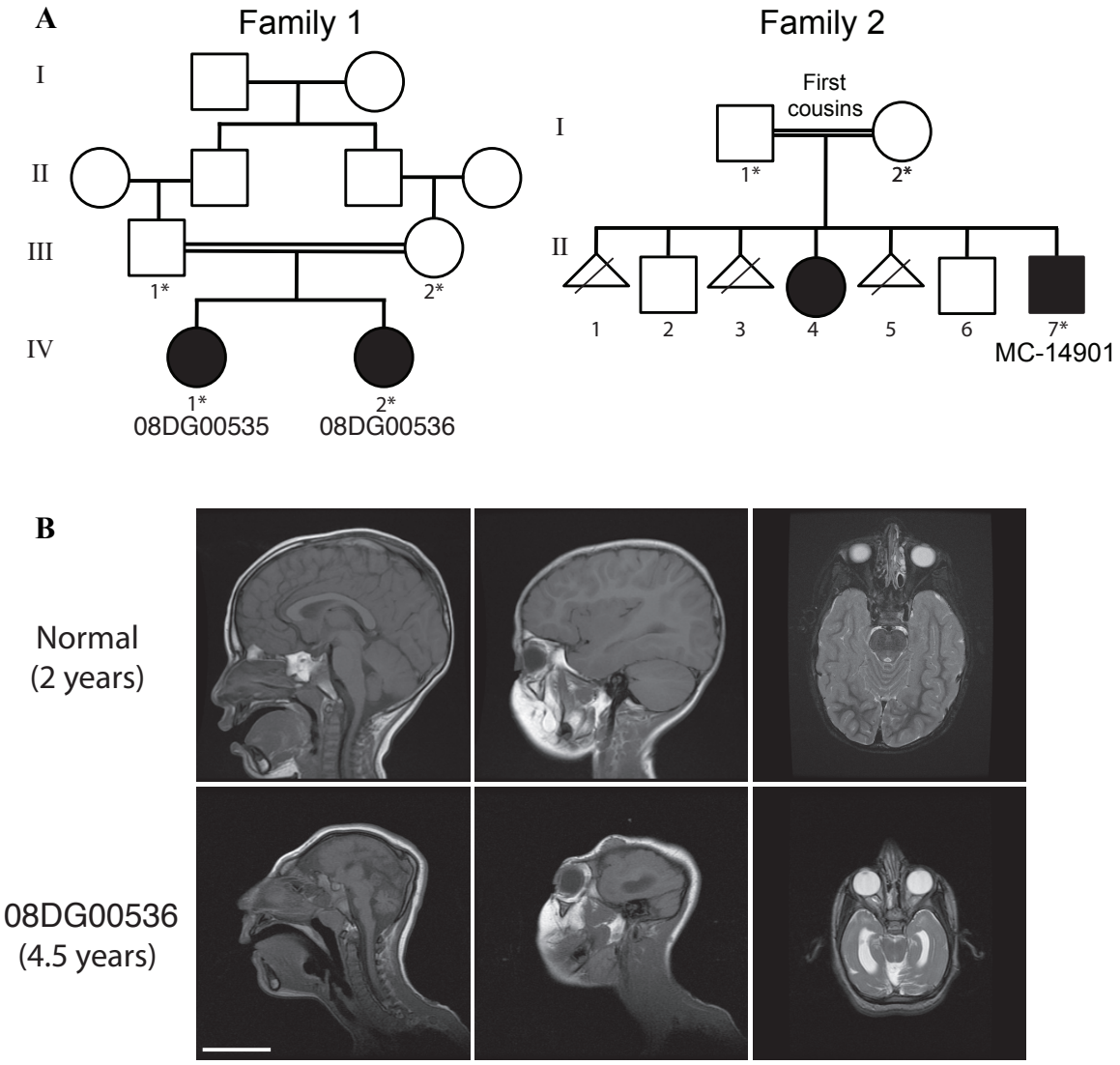


Figure 2-1. Pedigrees and radiographic features of the two consanguineous families with microlissencephaly.

(A) Both families are from Saudi Arabia. Parents of Family 1 are first cousins and have two affected female children (08DG00535, IV-1; 08DG00536, IV-2). Whole blood DNA from both parents and both affected children was obtained and analyzed (indicated by asterisk). Parents of Family 2 are first cousins with seven reported pregnancies, producing one affected male (MC-14901, II-7), one affected female (II-4) (not available for analysis), two unaffected males and three pregnancies that resulted in fetal demise. Whole blood DNA from both parents and the affected male (MC-14901, II-7) was obtained and analyzed (indicated by asterisk).

(B) Representative MRI images of 08DG00536 (family 1, IV-2) from Family 1 at 4.5 years of age, demonstrating the drastic reduction in brain size, agenesis of the corpus callosum, and abnormal gyral pattern compared to a normal 2 year old child. Sagittal T1 and axial T2 sections are shown (scale bar=5cm).

microcephaly with head circumference of 35 cm (13.4 standard deviations below the mean) and evidence of retarded growth with a length of 105 cm and weight of 10 kg (3.1 and 4.7 standard deviations below the mean respectively). She remained seizure-free but was only able to roll over, and her social and cognitive development was limited to spontaneous smiling. The second affected daughter (08DG00536; Family 1, IV-2 in **Figure 2-1a**) was first evaluated at birth, at which time her length and weight were normal, head circumference was 26.5 cm (5.6 standard deviations below the mean), and her anterior fontanel was almost closed. Her neurological course was static and characterized by marked hypertonia with virtually no gain of any milestone beyond spontaneous smiling and rolling over by the age of 5.5 years, but with no seizures. At that time her head circumference was 34 cm, length was 85 cm, and weight was 8.9 kg (12.7, 5.5 and 4.8 standard deviations below the mean respectively). Laboratory investigations including plasma acylcarnitines, carnitine, and amino acids, urine organic acids, ammonia, lactic acid, and high-resolution karyotype were all normal. MRI scans showed microcephaly, severe simplification of the gyral pattern (lissencephaly), agenesis of the corpus callosum, and colpocephaly (enlargement of the posterior lateral ventricles, often associated with corpus callosum defects), features that have been described in microlissencephaly (**Figure 2-1b**).

Family 2 originates from Western Saudi Arabia. The parents are healthy and are reported to be first cousins. They have a daughter and son with extreme microcephaly, two healthy sons, and had three additional pregnancies that resulted in spontaneous abortion (**Figure 2-1a**). The affected daughter was reported to have microcephaly without seizures but no additional information was available (Family 2, II-4 in **Figure 2-1a**). The

affected son (MC-14901; Family 2, II-7 in **Figure 2-1a**) displayed microcephaly and dysmorphic features at birth but growth parameters are not available. A head CT at that time revealed small brain size. Seizures began at two months of age and were described to start on the left side and progress to full-body convulsions with up-rolling of the eyes.

He was first evaluated at 9 months of age, at which time he could not roll or control his head and a neurologic exam revealed increased reflexes, decreased tone, normal power and positive Babinski signs. MR imaging at 11 months of age showed a marked decrease in the size of both cerebral hemispheres with a large midline fluid filled structure and dilatation of the right lateral ventricle, a small cerebellum, and agenesis of the corpus callosum. The thalami were not fused and a midline falx was noted. Laboratory investigations included a normal karyotype (46, XY) and an unremarkable acylcarnitine profile and amino acid analysis by tandem mass spectrometry. An evaluation at 3.5 years of age recorded his head circumference as 32 cm (11.1 standard deviations below the mean), length as 88 cm (2.8 standard deviations below the mean) and weight as 13 kg (6th percentile).

Identification of Homozygous NDE1 Mutations

All patients were born to consanguineous marriages, suggesting autosomal recessive inheritance and allowing homozygosity mapping, which identified only one common homozygous region larger than 1Mb--at chromosome 16p13.11--shared by all three affected children (DNA from the second affected child in Family 2 was unavailable). The maximal LOD scores of the two pedigrees at the homozygous region are 1.8 and 1.2, respectively. The homozygous locus is 4.6Mb in size and contains

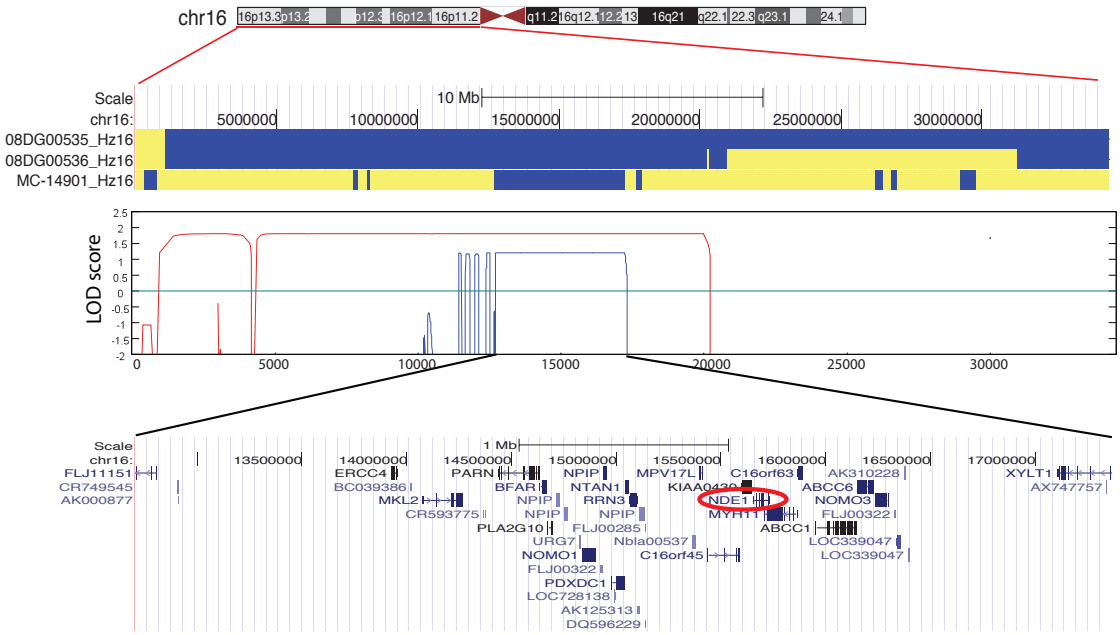


Figure 2-2. Homozygosity analysis of both pedigrees identified a 4.6-Mb region on chromosome 16 that is homozygous in all three affected children.
All homozygous SNPs are represented as blue and heterozygous
SNPs are represented as yellow. The maximal LOD scores of the two pedigrees at the homozygous region are 1.8 and 1.2, respectively. The shared region of homozygosity contains approximately 35 annotated genes, including *NDE1*.

approximately 35 annotated genes, including *NDE1* (**Figure 2-2**). *NDE1* was selected as the top candidate given that the mouse *Nde1* mutations strongly affected neural progenitor proliferation, and that defects of both proliferation and neuronal migration were observed in mice deficient for both Nde1 and Lis1 (Feng & Walsh 2004; Pawlisz et al. 2008; Yingling et al. 2008). Direct sequencing of NDE1 revealed frameshift mutations in both families (**Figure 2-3a,b**). Family 1 showed a c.684_685del mutation in exon 6 that creates a translational frameshift at codon 229 of the normal 335 amino acid protein sequence (p.Pro229TrpfsX85), and predicts a protein that consists of 312 amino acids, terminating after the addition of 84 abnormal amino acids (**Figure 2-3a,b**). Family 2 showed a c.733dup mutation in exon 7 that creates a frameshift mutation at codon 245 (into the same abnormal reading frame as the mutation in Family 1) creating a predicted protein truncated after the addition of 69 abnormal amino acids (p.Leu245ProfsX70).

Both mutations were homozygous in affected individuals, segregated perfectly with disease in each family, and were absent from more than 200 normal individuals, including 96 Saudi Arabian controls, and not seen in the 1000Genomes, strongly suggesting that they are bona fide mutations. This conclusion was further supported by identification of additional *NDE1* alleles in patients sharing similar clinical features by Bakircioglu et al. (2011).

Characterization of the Mutant NDE1 Proteins

Based on data from NDE1 and its paralogue, NDEL1, the truncated NDE1 proteins, if stable, would lack critical protein domains (**Figure 2-3a**). The C-terminus of

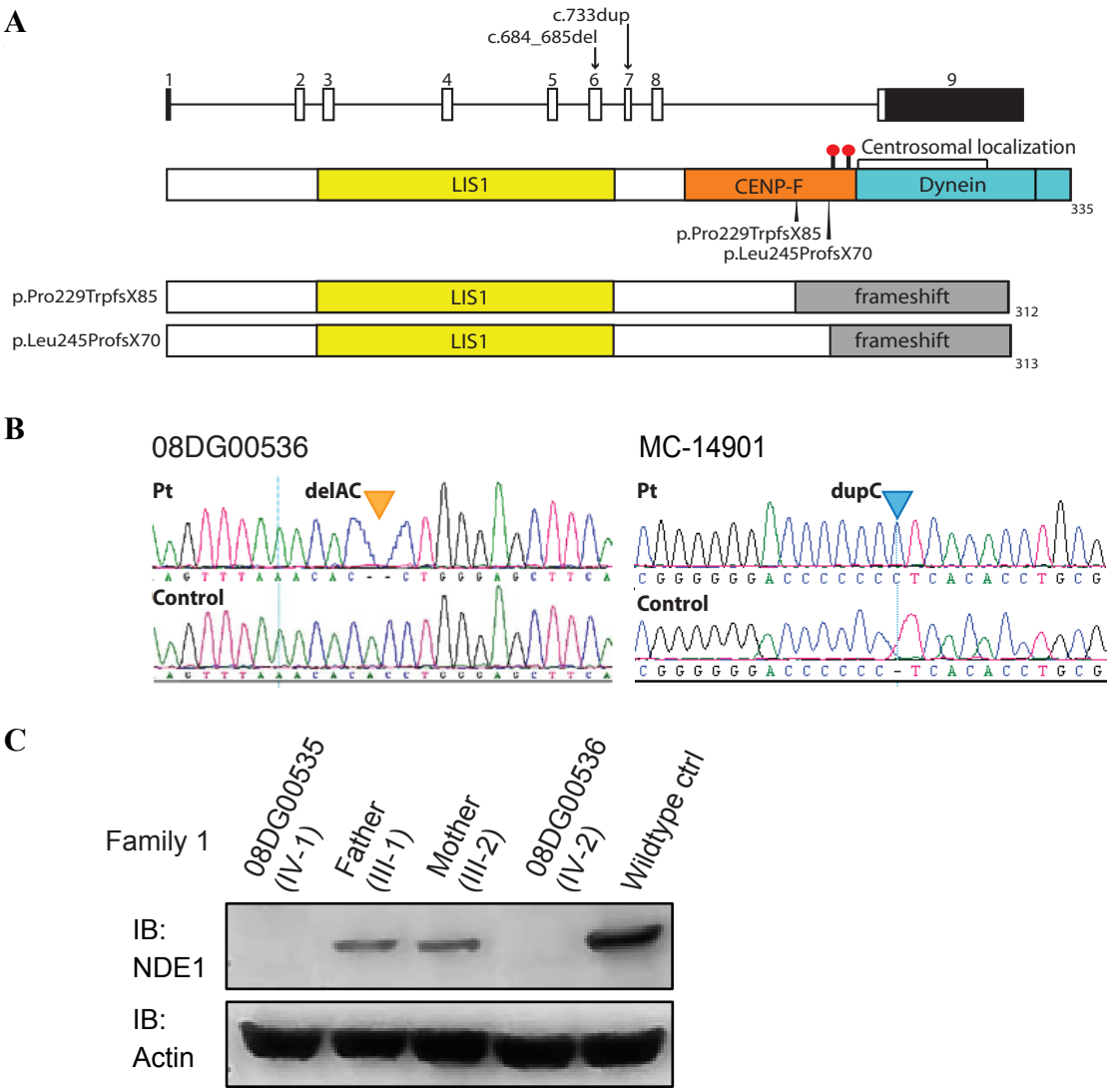


Figure 2-3. Two framshift homozygous mutations identified from NDE1.
(A) The human NDE1 gene consists of 9 exons (8 coding exons), which encode a protein with 335 amino acids, harboring multiple protein-interaction domains. Two frameshift mutations were identified in exon 6 and 7 respectively, both predicted to disrupt the CENP-F, dynein interaction domain, centrosomal localization domain, and at least two conserved phosphorylation residues implicated in mitotic progression (T246 and S250) at the C-terminus. The black bar shows noncoding exons; open bar, coding exon; red dot indicates potential phosphorylation sites.
(B) Representative Sanger sequencing traces indicating the two base-pair deletion identified in patient 08DG00536 (Family 1, IV-2 in **Figure 2-1a**) and the one base-pair duplicated identified in patient MC-14901 (Family 2, II-7 in **Figure 2-1a**.
(C) NDE1 protein was undetectable in the whole cell lysates collected from lymphoblasts of both patients (IV-1&2 in **Figure 2-1a**) in Family 1 harboring the c.684_685del (p.Pro229TrpfsX85) mutation. The protein levels of NDE1 from the parents (III-1&2 in **Figure 2-1a**) were reduced by roughly 50% compared to the wildtype control, consistent with their known heterozygous carrier status. Immunoblotting (IB) was performed with antibody against NDE1 and actin as a loading control.

NDE1 includes a domain required for interaction with CENP-F, which directs NDE1 to kinetochores (Vergnolle & Taylor 2007), and another domain that regulates binding to the cytoplasmic dynein complex and to Su4823, which was found recently to regulate centrosomal localization of NDE1 (Sasaki et al. 2000; Stehman et al. 2007; Hirohashi et al. 2006). Although RT-PCR using patient lymphoblasts from Family 1 suggested equal abundance of *NDE1* transcripts between patients and controls (data not shown), western blot analysis of patient lymphoblasts showed no detectable NDE1 protein in the two affected patients (Family 1, IV-I and IV-II) carrying the p.Pro229TrpfsX85 mutation (**Figure 2-3c**), while heterozygous parents showed ≈50% reduction in protein level. These data suggest that the frameshift mutation caused instability of the protein and subsequent degradation. Thus, the severe reduction of brain volume seen in these patients, which is far more severe than the dramatic *ASPM*-associated microcephaly (Bond et al. 2005; Bond et al. 2002; Desir et al. 2008), appears to be associated with loss of NDE1 protein, though some mutant protein in some tissues cannot be ruled out.

In order to further study the functional role of the C-terminus of NDE1, we engineered cDNAs corresponding to the p.Pro229TrpfsX85 and p.Leu245ProfsX70 alleles, tagged with FLAG (**Figure 2-3a**). Despite the instability of endogenous mutant proteins, overexpression allowed recovery of enough protein to study the effects of the two mutations on NDE1 binding to LIS1 and to dynein. Wild-type FLAG-NDE1 can be easily co-immunoprecipitated with the dynein complex (**Figure 2-4a**), and both frameshift mutants completely abolished dynein binding (**Figure 2-4b**). NDE1 also binds LIS1, but through the coiled-coil domain at the amino terminus of NDE1 (Feng et al. 2000; Derewenda et al. 2007),

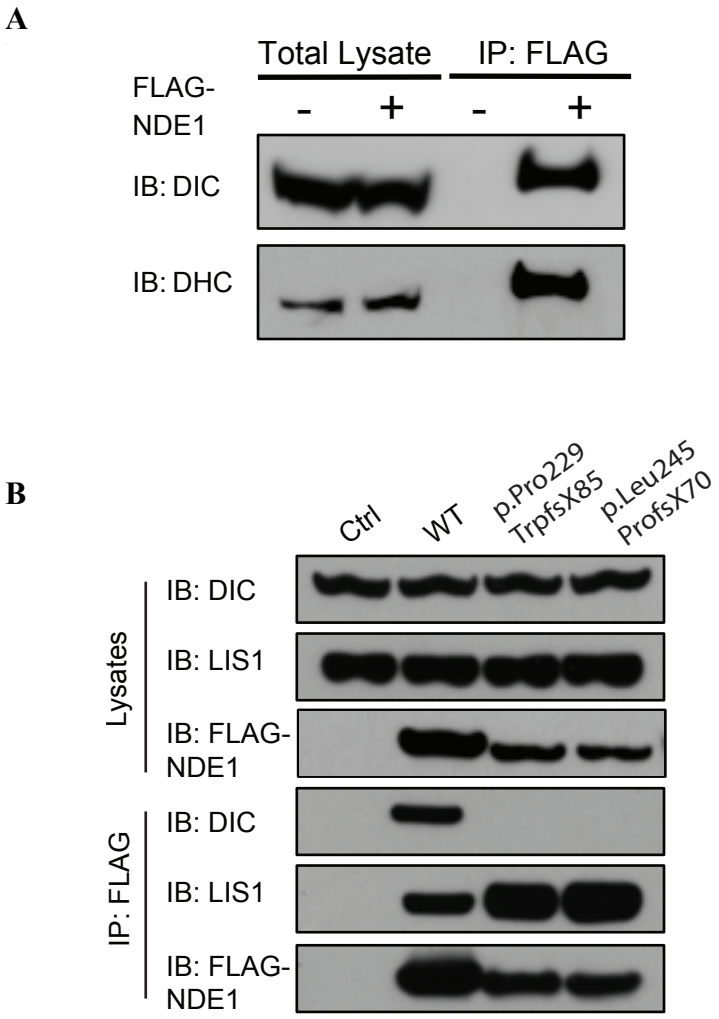


Figure 2-4. Both mutant alleles disrupt the interaction of NDE1 with the cytoplasmic dynein complex.

(A) WT FLAG-NDE1 interacts with the dynein complex. Immunoprecipitation was performed using anti-FLAG M2 beads. Cells transfected with empty 3XFLAG-CMV vector was labeled as "-" and used as the negative control; cells transfected with WT FLAG-NDE1 vector was labeled as "+". (DIC: dynein intermediate chain 74.1; DHC: dynein heavy chain)

(B) The interaction with dynein complex was abolished in both mutant NDE1 proteins, confirming the disruption of dynein-binding domain by both mutant alleles. However, the interaction with LIS1 was preserved or even slightly enhanced in both mutant proteins.

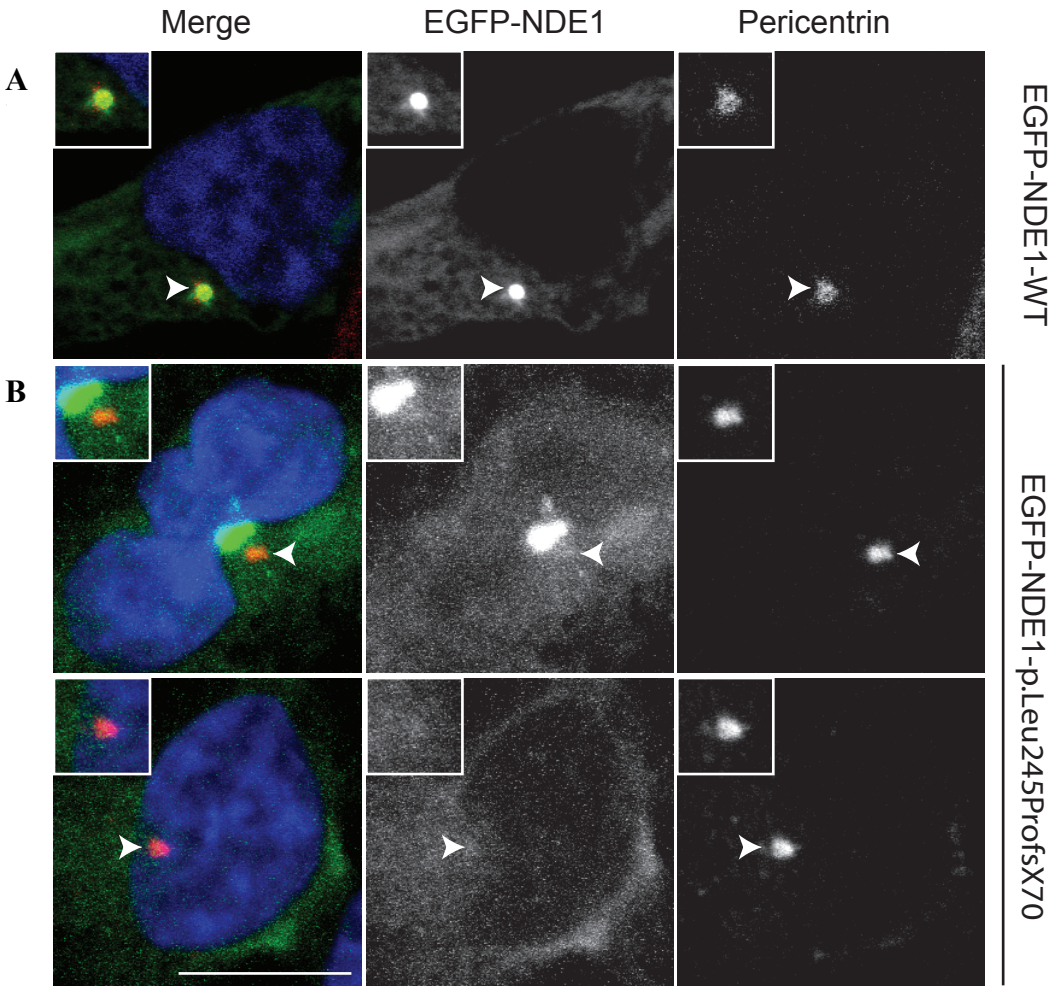


Figure 2-5. The centrosomal localization of NDE1 was abolished in the mutants proteins.
(A) Wild-type EGFP-NDE1-WT was transfected into 293T cells and localized to the centrosome. **(B)** Mutant EGFP-NDE1-p.Leu245ProfsX70 was transfected into 293T cells and did not localize to the centrosome. It either formed non-centrosomal aggregates, or localized diffusely in the cytoplasm. Centrosomes labeled by Pericentrin are indicated by the arrowhead and magnified in the upper right inlets (scale bar=10μm).

and LIS1 binding was normal or even enhanced (**Figure 2-4b**) in the FLAG-tagged mutant proteins. Since the cytoplasmic dynein complex has profound roles in neurogenesis, including mitotic spindle organization, interkinetic nuclear migration, and neuronal migration, the loss of an important dynein regulator could impact multiple aspects of neurogenesis.

Both mutations are also predicted to abolish the centrosomal localization domain located at the C-terminus of NDE1 (**Figure 2-3a**). To examine the effects of mutations on NDE1 subcellular localization, GFP tagged wild-type or p.Leu245ProfsX70 mutant were transfected in 293T cells. Wildtype GFP-NDE1 localized to the centrosome, consistent with previous findings (**Figure 2-5a**), whereas mutant GFP-NDE1-p.Leu245ProfsX70 (c.733dup) failed to target the centrosome but either presented as non-centrosomal aggregates or diffusely in the cytoplasm. (**Figure 2-5b**). Similar results were reported by analyzing the p.Pro229TrpfsX85 (c.684_685del) allele by Bakircioglu et al. (2011). Therefore both NDE1 mutations disrupt at least two key functions of NDE1, suggesting that the developmental defects seen in the patients are likely caused by the loss of NDE1 function.

Loss of NDE1/Nde1 Disrupts Mitotic Progression in Both Human and Mice

Mouse embryonic fibroblasts with Nde1 mutations (**Figure 2-6**) showed defects in mitotic progression evident by increased mitotic index despite growing more slowly in culture (**Figure 2-6a**), and abnormal spindle structures such as multipolar spindles and chromosome misalignment (**Figure 2-6b**). Similarly, patient-derived lymphoblast cells (**Figure 2-7**) also show defects in spindle structure, including tripolar spindles, misaligned mitotic chromosomes, nuclear fragmentation, and abnormal microtubule

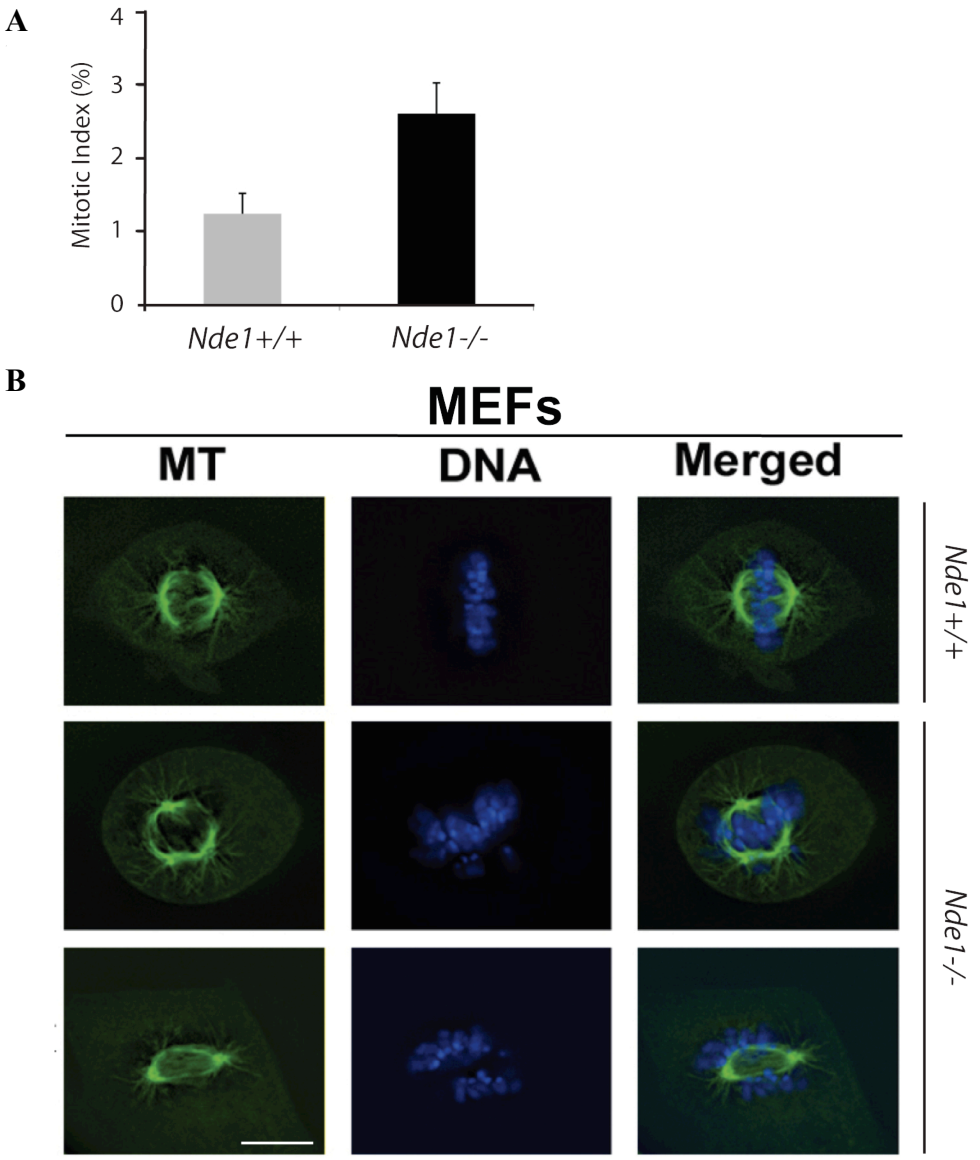


Figure 2-6. *Nde1* deficiency leads to increased mitotic index and abnormal mitotic spindles in early passage (P0-3) primary MEFs.

(A) Mitotic index increased by approximately 50% in *Nde1*$^{-/-}$ MEFs compared to *Nde1*$^{+/+}$ MEFs at passage 1 (p-value <.01, Chi-square test of homogeneity for two independent samples).

(B) Primary MEFs derived from wild type (*Nde1*$^{+/+}$) and mutant (*Nde1*$^{-/-}$) embryos were analyzed directly for the structure of mitotic spindles by staining with monoclonal antibody to tubulin (in green) and Hoechst for chromosomal DNA (in blue). Examples of normal *Nde1*$^{+/+}$ M-phase cell, an abnormal *Nde1*$^{-/-}$ M-phase cell with tripolar mitotic spindle and mis-aligned mitotic chromosomes, and an abnormal *Nde1*$^{-/-}$ mitotic cell with discordant mitotic spindle and chromosome alignments are shown (scale bar=10μm).

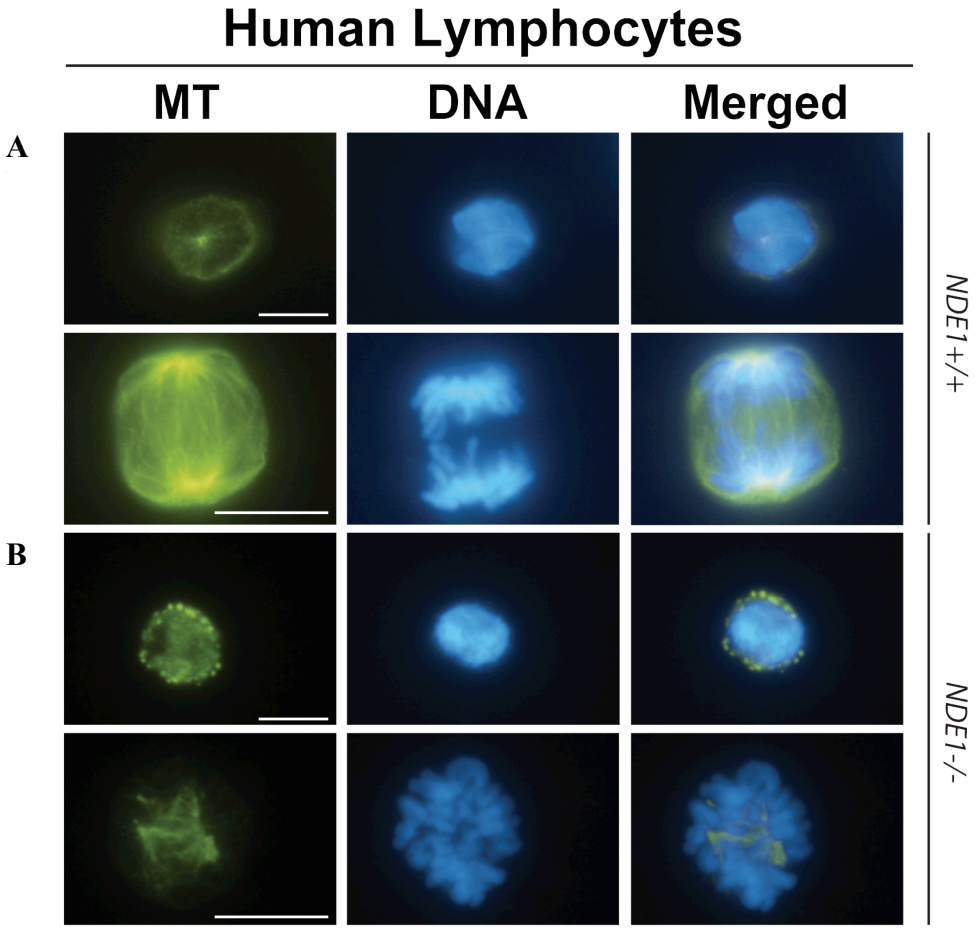


Figure 2-7. *NDE1* patient lymphoblasts exhibit defects in mitotic spindle organization.
(A) Control lymphoblasts showed normal looking nuclei and normal alpha-tubulin staining. Examples of a $NDE1^{+/+}$ interphase cell and a $NDE1^{+/+}$ M-phase cell are shown.
(B) Patient lymphoblasts showed condensed/fragmentized nuclei and disorganized alpha-tubulin. g, Examples of an abnormal $NDE1^{-/-}$ interphase cell with nuclear fragmentation and an abnormal $NDE1^{-/-}$ M-phase cell with multipolar and disorganized spindles (scale bar=10μm) are shown. The percentage of mitotic cells found with abnormal spindles was 7% in patient lymphoblasts, compared to 0% found in control lymphoblasts ($P<0.02$, two-tailed Chi-squared test).

organization, further supporting that NDE1's is essential role in normal mitotic spindle function. Patient lymphoblast cultures also showed excessive dying cells (not shown), which could also contribute to the developmental defects.

Discussion

Our data indicate that loss of NDE1 produces profound defects in cerebral cortical size and organization, as well as less profound defects in somatic size. Both of the identified *NDE1* mutations truncate the C terminus of the NDE1 protein, preventing normal dynein binding. Whereas the overexpressed proteins are somewhat unstable, they retain LIS1 binding activity, and so in principle could act as dominant-negative proteins. However, the absence of a genetically dominant phenotype, the absence of detectable mutant protein in patient cells, and the similarity of phenotype in the three patients described here (as well as additional patients and an additional mutant allele described by Bakircioglu et al., 2011) all argue strongly that the mutations act as simple null alleles.

The C-terminus of *NDE1* is also mutated by the recurrent inversions of 16p that are associated with acute myelogenous leukemia (AML), which disrupts exon 8 of *NDE1*, encoding the extreme C-terminus of the protein (Van der Reijden et al. 2010). Existing pedigree analysis and medical follow-up of Families 1 and 2, while somewhat limited, has so far not suggested the presence of AML or other leukemia in homozygous or heterozygous carriers of *NDE1* mutations, though more extensive analysis of additional families is needed. On the other hand, our biochemical analysis suggests that the AML disruptions, if they create a stable truncated protein, could potentially create abnormal

proteins with preserved LIS1 binding but abnormal dynein binding, which might contribute to leukemogenesis.

Although mice lacking *Nde1* show about a one-third reduction in brain mass (Feng & Walsh 2004), neuronal migration is only moderately deficient, whereas human NDE1 mutations cause >50% decrease in cortical volume, and striking architectural disturbances that suggest severely abnormal neuronal migration. The more striking brain defects seen in humans harboring *NDE1* mutations, especially the marked architectural defects, could reflect morphological or quantitative defects in the radial glial cells in human *NDE1* deficiency that normally act as guides for migrating neurons, or may merely reflect the larger human brain with more profound defects in neurogenesis. Alternatively, the differences could reflect a greater role for *NDE1* in migrating neurons in humans than in mice, parallel to the much more profound defects in neuronal migration in humans versus mice after removal of one allele of LIS1. The Nde1 paralogue, Ndel1, is highly similar structurally to Nde1, more highly expressed in post-mitotic neurons, and appears to have larger roles in cell proliferation outside the brain, mitotic spindle orientation, and neuronal migration (Wynshaw-Boris 2007; Derewenda et al. 2007; Mori et al. 2007; Shu et al. 2004; Lam et al. 2010; Sasaki et al. 2005; Toyo-Oka et al. 2005; Yamada et al. 2008). The details of potential genetic redundancy between NDE1 and NDEL1 may also contribute to these mouse-human differences.

Humans with NDE1 mutations show modestly reduced height and weight as well as cerebral cortical size, though the defect in head circumference (typically -10 to -14 SD below the mean) was more marked statistically than defects in height and weight (typically -2 to -5 SD below mean). The decreased body size could reflect roles of NDE1

in other tissues, although more specific hypothalamic defects, or nutritional explanations related to the patients' poor neurological function, cannot be ruled out. Mice with Nde1 mutations show at most a slight, statistically insignificant, reduction in body mass (Feng & Walsh 2004). Since *NDE1*, like many other microcephaly genes, is expressed in many developing tissues (Feng et al. 2000), the more severe involvement of brain is generally regarded as reflecting the more limited ability of the brain to regulate its size, given that most brain cells are post-mitotic, but other mechanisms may also contribute to this tissue specificity.

Finally, although analysis of archived DNA samples from the original microlissencephaly family of Norman and Roberts (Norman et al. 1976) did not reveal a detectable mutation in *NDE1* (data not shown), the similarity of the *NDE1* phenotype to the "Norman-Roberts" syndrome is striking, and some other cases of microlissencephaly also do not show NDE1 mutations. Hence, this classical form of microlissencephaly may ultimately reflect defects in proteins of this highly conserved centrosomal pathway that function in close concert with NDE1.

Since the manuscript was published in 2011, an additional mutant allele of *NDE1* have been identified, leading to a slightly different neurodevelopmental disorder, named microhydranencephaly, characterized by extreme microcephaly with ventricular dilatation but without obvious lissencephaly (Guven et al. 2012). This additional allele was identified from a Turkish family with a 4.3kb homozygous deletion that encompasses the start codon in exon 2, and therefore is predicted to be a null allele (Guven et al. 2012). Its identification widens the spectrum of brain malformations caused by *NDE1* mutations, highlighting a multifunctional role of NDE1 in brain development.

Although we are accustomed to think that alternate alleles within a given gene are likely to cause similar phenotypes, more and more cases studied recently have started to challenge this paradigm. In addition to the example of *NDE1* we observed here, brain malformations caused by *WDR62* (i.e. the MCPH2 gene) mutations showed broad phenotypic diversity involving both neurogenesis and neuronal migration defects, suggesting potential functional convergence of these two cellular processes (Nicholas et al. 2010; Yu et al. 2010; Bilguvar et al. 2010). In fact, mutations in the physical interaction partners of NDE1, such as LIS1, dynein heavy chain (DHC) and katanin p80 (unpublished data from Walsh lab) also seem to be involved in a wide range of cellular processes, and have all been identified to cause related brain malformations, further suggesting that genetic interactions can serve as a mean to map molecular regulatory networks for key developmental processes such as neurogenesis and neuronal migration (Walsh & Engle 2010). More interestingly, both LIS1 and DHC mutations are identified as single-hit *de novo* mutations, consistent with their critical roles in multiple fundamental cellular functions ranging from spindle formation during mitosis, cytokinesis, cellular polarity, nucleokinesis and leading process extension.

Materials and Methods

Human studies

The human studies protocols were reviewed and approved by the institutional review boards of the Children's Hospital Boston and the King Faisal Specialist Hospital and Research Centers at Riyadh and Jeddah, and human research was performed in

accordance with the ethical standards with proper informed consent obtained. Standard protocols were used for blood draw and DNA extractions.

Genome-wide linkage analysis

Family 1 was genotyped using the Affy 250K StyI SNP Chip as per the manufacturer's protocol. Family 2 was genotyped using the Illumina660W-Quad Chip at the W.M. Keck Foundation Biotechnology Resource Laboratory at Yale University. Single and multipoint LOD scores were calculated using Allegro assuming a recessive mode of disease inheritance, full penetrance, and a disease allele frequency of 0.0001. Nucleotide numbers are in reference to cDNA (RefSeq NM_017668.2, where A of the ATG translational start site is designated as +1) coordinates, and amino acid numbers are in reference to protein (RefSeq NP_001137451.1) coordinates, all following HGVS guidelines.

Sanger sequencing

Coding *NDE1* exons as well as flanking intronic sequences were amplified by PCR, followed by bidirectional sequencing using ABI 3730XL DNA Analyzer or submitted to Polymorphic DNA Technologies for Sanger capillary electrophoresis. Sequencing of >200 neurologically normal control samples and 96 unaffected individuals from Saudi Arabian families with unrelated disorders failed to identify either mutant variant.

Cell culture, transfection, cell synchronization and flow cytometry

293T cells were cultured in DMEM containing 10% fetal bovine serum (FBS). Mouse embryonic fibroblasts (MEFs) were isolated from E13.5 *Nde1*$^{-/-}$ mice and their littermates, and were cultured in Dulbecco's Modified Earle Medium (DMEM) with 10%

FBS for 2 passages or fewer. Human lymphoblasts from normal individuals, *NDE1*$^{-/-}$ individuals, and heterozygous *NDE1*$^{-/+}$ parents were transformed using Epstein-Barr virus and cultured in RPMI medium supplemented with 15% fetal calf serum (FCS). For FLAG-NDE1 overexpression, 293T cells were transfected with specified plasmids (empty FLAG-containing vector, FLAG-NDE1-WT, FLAG-NDE1-delAC and FLAGNDE1-dupC) using Fugene6® (Roche).

Western blotting and immunoprecipitation

Cell lysates were prepared either using lysis buffer (for immunoprecipitation) containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 0.4% NP-40, 1 mM NaF, 10 mM b-glycero-phosphate, 10 nM calyculin A, 1 mM Na3VO4, 1 mM PMSF and protease inhibitor cocktail mix (Roche) or RIPA buffer (50 mM Tris-HCl 7.4, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS) containing protease inhibitor cocktail mix (Roche) for direct western blotting. Protein concentrations were normalized using the BCA assay (Thermo Scientific). For immunoprecipitation, cell lysates were incubated with anti-FLAG M2 affinity gel (Sigma) for 2 hours at 4°C, followed by three-times wash using the same lysis buffer. Protein samples were then eluted using 3XFLAG peptides (150 ng/μl) for 1 hour at 4°C. For western blotting, cell lysates or immunoprecipitation elution were boiled in LDS sample buffer (Invitrogen), followed by electrophoresis on 4-12% Bis-Tris or 7% Tris-Acetate SDS-PAGE (Invitrogen) and transfer onto PVDF membrane (Millipore). Membranes were blocked in TBST containing 5% non-fat milk or Odyssey Blocking Buffer (LICOR Biosciences) at room temperature for 30 minutes, and incubated with primary antibodies according to the antibody manufacture's instruction, followed by incubation with HRP-conjugated

secondary antibodies (Cell Signaling Technology) or fluorescent-dye-conjugated secondary antibodies (LICOR Biosciences). Immunosignals were detected by SuperSignal West Pico Chemiluminescent (Pierce) or the Odyssey® Infrared Imaging System (LICOR Biosciences).

The rabbit antibody against total Nde1 was used as previously described (Feng & Walsh 2004). In addition, the following antibodies were purchased and used according to the manufacturer's instructions: mouse anti-FLAG (M2) antibody and beads (Sigma); rabbit anti-DYKDDDDK (FLAG) (Cell Signaling Technology); rabbit anti-LIS1 (Bethyl Laboratory); mouse monoclonal anti-dynein intermediate chain (clone 74.1) (Millipore); rabbit anti-dynein heavy chain (Santa Cruz Biotechnology); mouse monoclonal anti-Cdk1 (Millipore).

For the FLAG-NDE1 construct, human *NDE1* cDNA (RefSeq NM_017668.2, starting from ATG) was PCR amplified and subcloned into the p3XFLAG-CMV-10 vector (Sigma); the two mutant constructs were generated by site-directed mutagenesis to delete 684_685AC (c.684_685del) or duplicate 733C (c.733dup), respectively.

References

Bakircioglu, M. et al., 2011. The essential role of centrosomal NDE1 in human cerebral cortex neurogenesis. *American journal of human genetics*, 88(5), pp.523–535.

Bilguvar, K. et al., 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, 467(7312), pp.207–210.

Bond, J. et al., 2005. A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nature Genetics*, 37(4), pp.353–355.

Bond, J. et al., 2002. ASPM is a major determinant of cerebral cortical size. *Nature Genetics*, 32(2), pp.316–320.

Derewenda, U. et al., 2007. The structure of the coiled-coil domain of Ndel1 and the basis of its interaction with Lis1, the causal protein of Miller-Dieker lissencephaly. *Structure (London, England : 1993)*,

15(11), pp.1467–1481.

Desir, J., Cassart, M. & David, P., 2008. Primary microcephaly with ASPM mutation shows simplified cortical gyration with antero-posterior gradient pre-and post-natally. *American Journal of …*.

Dobyns, W.B. et al., 1993. Lissencephaly. A human brain malformation associated with deletion of the LIS1 gene located at chromosome 17p13. *JAMA : the journal of the American Medical Association*, 270(23), pp.2838–2842.

Dobyns, W.B., Stratton, R.F. & Greenberg, F., 1984. Syndromes with lissencephaly. I: Miller-Dieker and Norman-Roberts syndromes and isolated lissencephaly. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 18(3), pp.509–526.

Efimov, V.P. & Morris, N.R., 2000. The LIS1-related NUDF protein of Aspergillus nidulans interacts with the coiled-coil domain of the NUDE/RO11 protein. *The Journal of cell biology*, 150(3), pp.681–688.

Feng, Y. & Walsh, C.A., 2004. Mitotic spindle regulation by Nde1 controls cerebral cortical size. *Neuron*, 44(2), pp.279–293.

Feng, Y. et al., 2000. LIS1 regulates CNS lamination by interacting with mNudE, a central component of the centrosome. *Neuron*, 28(3), pp.665–679.

Gleeson, J.G. et al., 1998. Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. *Cell*, 92(1), pp.63–72.

Guven, A. et al., 2012. Novel NDE1 homozygous mutation resulting in microhydranencephaly and not microlyssencephaly. *Neurogenetics*, 13(3), pp.189–194.

Hirohashi, Y. et al., 2006. Centrosomal proteins Nde1 and Su48 form a complex regulated by phosphorylation. *Oncogene*, 25(45), pp.6048–6055.

Hong, S.E. et al., 2000. Autosomal recessive lissencephaly with cerebellar hypoplasia is associated with human RELN mutations. *Nature Genetics*, 26(1), pp.93–96.

Lam, C. et al., 2010. Functional interplay between LIS1, NDE1 and NDEL1 in dynein-dependent organelle positioning. *Journal of cell science*, 123(Pt 2), pp.202–212.

McKenney, R.J. et al., 2010. LIS1 and NudE induce a persistent dynein force-producing state. *Cell*, 141(2), pp.304–314.

Mori, D. et al., 2007. NDEL1 phosphorylation by Aurora-A kinase is essential for centrosomal maturation, separation, and TACC3 recruitment. *Molecular and cellular biology*, 27(1), pp.352–367.

Nicholas, A.K. et al., 2010. WDR62 is associated with the spindle pole and is mutated in human microcephaly. *Nature Genetics*, 42(11), pp.1010–1014.

Norman, M.G. et al., 1976. Lissencephaly. *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*, 3(1), pp.39–46.

Pawlisz, A.S. et al., 2008. Lis1-Nde1-dependent neuronal fate control determines cerebral cortical size and lamination. *Human molecular genetics*, 17(16), pp.2441–2455.

Poirier, K. et al., 2007. Large spectrum of lissencephaly and pachygyria phenotypes resulting from de novo missense mutations in tubulin alpha 1A (TUBA1A). *Human mutation*, 28(11), pp.1055–1064.

Sasaki, S. et al., 2000. A LIS1/NUDEL/cytoplasmic dynein heavy chain complex in the developing and adult nervous system. *Neuron*, 28(3), pp.681–696.

Sasaki, S. et al., 2005. Complete loss of Ndel1 results in neuronal migration defects and early embryonic lethality. *Molecular and cellular biology*, 25(17), pp.7812–7827.

Shu, T. et al., 2004. Ndel1 operates in a common pathway with LIS1 and cytoplasmic dynein to regulate cortical neuronal positioning. *Neuron*, 44(2), pp.263–277.

Stehman, S.A. et al., 2007. NudE and NudEL are required for mitotic progression and are involved in dynein recruitment to kinetochores. *The Journal of cell biology*, 178(4), pp.583–594.

Thornton, G.K. & Woods, C.G., 2009. Primary microcephaly: do all roads lead to Rome? *Cell*, 25(11), pp.501–510.

Toyo-Oka, K. et al., 2005. Recruitment of katanin p60 by phosphorylated NDEL1, an LIS1 interacting protein, is essential for mitotic cell division and neuronal migration. *Human molecular genetics*, 14(21), pp.3113–3128.

Van der Reijden, B.A. et al., 2010. The NDE1 gene is disrupted by the inv (16) in 90% of cases with CBFB–MYH11-positive acute myeloid leukemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*.

Vergnolle, M.A.S. & Taylor, S.S., 2007. Cenp-F links kinetochores to Ndel1/Nde1/Lis1/dynein microtubule motor complexes. *Current biology : CB*, 17(13), pp.1173–1179.

Walsh, C.A. & Engle, E.C., 2010. Allelic diversity in human developmental neurogenetics: insights into biology and disease. *Neuron*, 68(2), pp.245–253.

Wynshaw-Boris, A., 2007. Lissencephaly and LIS1: insights into the molecular mechanisms of neuronal migration and development. *Clinical genetics*, 72(4), pp.296–304.

Wynshaw-Boris, A. et al., 2010. Lissencephaly: mechanistic insights from animal models and potential therapeutic strategies. *Seminars in cell & developmental biology*, 21(8), pp.823–830.

Yamada, M. et al., 2008. LIS1 and NDEL1 coordinate the plus-end-directed transport of cytoplasmic dynein. *The EMBO journal*, 27(19), pp.2471–2483.

Yingling, J. et al., 2008. Neuroepithelial stem cell proliferation requires LIS1 for precise spindle orientation and symmetric division. *Cell*, 132(3), pp.474–486.

Yu, T.W. et al., 2010. Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. *Nature Genetics*, 42(11), pp.1015–1020.

Chapter 3: Whole-genome amplification of single neurons from human brain and identification of somatic L1 insertions

This chapter contains work from the manuscript "Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain", published in *Cell*, October 26, 2012; 151(3):483-496. The text and figures were modified to fit the format of this dissertation. Xuyu Cai was co-first author on this manuscript with Gilad Evrony, and all experiments and data analyses in this chapter were carried out jointly by Gilad Evrony and Xuyu Cai.

Summary

This chapter summarizes the methods development of single-cell genomic sequencing from fresh frozen human brain tissues. Single neuronal and non-neuronal nuclei, from postmortem or surgically removed brain tissues, were obtained by successive nucleus purification, immunostaining and fluorescence-activated cell sorting (FACS). Two independent whole-genome amplification methods were applied to the isolated nuclei to amplify genomes from single cells for downstream analyses. Extensive quality controls were performed to ensured sorting and amplification success, as well as assessing the genotype concordance and genome coverage at various scales. The development of such methods allows us to systematically assess the genomic variance for a wide spectrum of somatic mutation types between single neurons from human brains. Potential future applications and limitations of the methods are also discussed.

Introduction

Isolation and amplification of single neuronal genome has been technically challenging due to the extremely small DNA content of a single cell (~6pg). No current sequencing technology is sensitive enough to sequence such a small amount of input material; therefore, extensive whole-genome amplification is required. Whole-genome amplification is known to introduce technical artifacts that complicate the analyses; therefore, a careful survey and comparison of different cell isolation and whole-genome amplification technique was carried out prior to the launch of the project.

Isolation of single cells

The first challenge of the study is to isolate a large number of single cells of the cell type of interest (e.g. neurons vs. glia) free of external contamination. The methods available for single cell isolation include: fluorescence-activated cell sorting (FACS), microfluidics, laser-capture microdissection (LCM) and related robotic and manual micro-manipulation tools (Navin & Hicks 2011). Each of these methods has their pros and cons, which were taken into consideration before choosing the most suitable method for our purpose. Among them, FACS is the most widely used technology for cell type isolation based on fluorescence labeling, particularly in the field of immunology. It also has the highest throughput allowing the sorting thousands of cells within minutes. A number of studies have successfully applied FACS to isolate neuronal versus non-neuronal nuclei based on immunoreactivity to the pan-neuronal marker NeuN (Spalding et al. 2005; Rehen et al. 2005; Matevossian & Akbarian 2008). However, FACS technology was primarily designed for bulk sorting; therefore, although it's capable of single cell sorting, its ability to consistently deposit the desired small number of cells into 96-well or 384-well PCR plates for downstream processing remained questionable.

LCM and related techniques allow isolation of single cells from thin-sectioned tissue specimens. Compared to FACS, LCM not only allows the identification of desired cell types based on immunofluorescence labeling, but also on the histology. This represents a big advantage in the isolation of rare cell populations that might lack reliable immunofluorescence markers. On the other hand, LCM requires extensive tissue manipulations including tissue sectioning, fixation and staining prior to the procedure. These procedures significantly lengthen the preparation time and pose risks to the genome integrity of processed tissues. For instance, tissue fixation with aldehydes (e.g.,

formalin, parafaormaldehyde) is known to denature, modify, and fragment the genomic DNA and thereby affects the downstream whole-genome amplification. Additionally, sectioning tissue into 10-20μm slices may destroy the nuclear integrity of targeted cells, resulting in incomplete recovery of their genome. Moreover, LCM is a low-throughput method that takes hours to process up to 10-20 cells.

Microfluidic systems are an emerging technology that allows accurate isolation of single cells or even single chromosomes into small chambers in nanoliter size (Zare & Kim 2010; Fan et al. 2011). Small volumes of reaction cocktail can be subsequent flowed into the chambers for whole-genome amplification. Such a setup provides significant advantages over the other two technologies in preventing external contamination, and this has been critical for single bacterial genomics studies (Blainey & Quake 2011; Zhang et al. 2006). Moreover, the small nanoliter reaction volume appears to improve amplification bias compared to standard microliter reactions, as well as greatly reducing the reagent costs for large-scale studies (Marcy et al. 2007). Despite all the advantages, isolation of fluorescence-labeled cells can be tricky with microfluidic systems as it requires manual calls on whether to include a cell under the microscope. In contrast, FACS can analyze tens of thousands cells at a time to build statistical signal intensity distributions for clearer separation of positive versus negative cells. Moreover, microfluidic systems are difficult to set up and require expertise in usage and maintenance. Only recently, bench-top microfluidic systems for single cell analysis became commercially available (e.g. Fluidigm C1 Single-cell system), which are expected to significantly improve the accessibility of the technology to non-expert end

users. However, to date, the system only accommodates single cell RNA amplification instead of whole-genome DNA amplification.

After considering the above concerns, we decided to use FACS to isolate single neurons given its high throughput, accessibility, and previous successful examples of isolating neuronal nuclei (Spalding et al. 2005; Matevossian & Akbarian 2008). Meanwhile, we are aware of its technical limitations, including its questionable accuracy on sorting single cells and potential risk of introducing external contamination during the process. These concerns were carefully evaluated while developing the methods.

Single cell whole-genome amplification

The second challenge of this study is to consistently amplify single genomes for downstream analyses. A variety of technical artifacts and biases can be introduced by the currently available whole-genome amplification; these potential artifacts need to be carefully evaluated and considered in the development of methods. There are two major methods for whole-genome amplification from single cells.

The first method, named multiple displacement amplification (MDA), is an isothermal reaction using a phi29 polymerase, which has a strong displacement activity that allows the polymerase to displace double-strand DNA to continuously create free single-strand DNA for further reaction at the same temperature. The reaction starts with alkaline lysis and denaturation of the original template, followed by random hexamer annealing and amplification at 30C (Dean et al. 2002). In additional to its displacement activity, phi29 shows a robust proof-reading activity that limits the error rate down to 10^{-7}, as well as its great processivity which leads to the generation of overlapping amplicons up to ~30kb in size (Hou et al. 2012). Recent studies showed that single cancer cells can

be amplified by MDA to recover up to 90% of the genome in a fashion that was suitable for studies of single-nucleotide variants (SNV) via whole-genome and whole-exome sequencing (Hou et al. 2012; Xu et al. 2012). However, MDA is known to suffer from non-linear amplification such that different regions of the genome stochastically get under or over-amplified (Dean et al. 2002). Additionally, a small fraction of the genome is consistently under-amplified or completely "dropped out" from MDA presumably due to high GC content or secondary structures of these regions (Evrony et al. 2012). The non-linear amplification and regional dropouts are expected to complicate the analysis of copy number variants (CNVs) from the amplified single cells. Lastly, MDA is known to create chimera sequences due to polymerase slippage and branch migration; these chimeras can be falsely classified as structural variants (SVs) during downstream analyses.

A second amplification method, named GenomePlex WGA4 (licensed by Sigma, later referred as "WGA4"), is a PCR-based method that starts with denaturation and fragmentation of the single cell genome, followed by universal adaptor ligation and PCR amplification. Since the genome is fragmented into ~400bp fragments prior to amplification, the amplification is generally more linear and produces higher quality copy number profiles from amplified single cells. This approach was successfully applied to the first human single cell study, which performed copy number profiling via low coverage sequencing to study the evolution of breast cancer cells (Navin et al. 2011). However, due to the initial fragmentation of the genome, only ~10% of the single cell genome is recovered by this method. Moreover, the fragmentation results in non-overlapping amplicons, which makes PCR-based secondary validation impossible. Taken

together, although WGA4 is a better method for CNV analysis, MDA is required for study other types of genetic variants, including SNVs, structural variants (SV) and retrotransposition. Both methods were tested in this chapter and their performance and applications will be discussed later in the chapter.

Results

Isolation of single neuronal nuclei from human brain tissue

We purified nuclei from post-mortem human frontal cortex and caudate nucleus and labeled them with a neuron-specific antibody (NeuN) for sorting using fluorescence-activated cell sorting (FACS) (**Figure 3-1A**) (Matevossian & Akbarian 2008; Spalding et al. 2005). Large nuclei with neuronal nuclear morphology were readily apparent by microscopy (**Figure 3-1B**) (Parent & Carpenter 1995). NeuN immunoreactivity (**Figure 3-1C**) labels essentially all neuronal nuclei in cortex and caudate, corresponding to 25-35% of all nuclei (population I; **Figures 3-2**) (Mullen et al. 1992; Wolf et al. 1996). Consistent with their increased size on microscopy (**Figure 3-1B**), $NeuN^+$ nuclei also had larger forward (FSC) and side (SSC) scatter (correlates of size) by flow cytometry compared to $NeuN^-$ nuclei (**Figures 3-2C**). Whereas for nuclei isolated from the caudate we performed a simple sort of the $NeuN^+$ population (population I, **Figures 3-2B**), we further enriched nuclei from the cortex for pyramidal neuronal nuclei. Since neighboring cortical pyramidal neurons tend to have shared clonal origins due to their primarily radial migration (Magavi et al. 2012), enriching for pyramidal neuronal nuclei increases the chance of identifying clonal somatic mutations shared by multiple neurons. The largest neuronal nuclei in cortex correspond primarily to pyramidal projection neurons

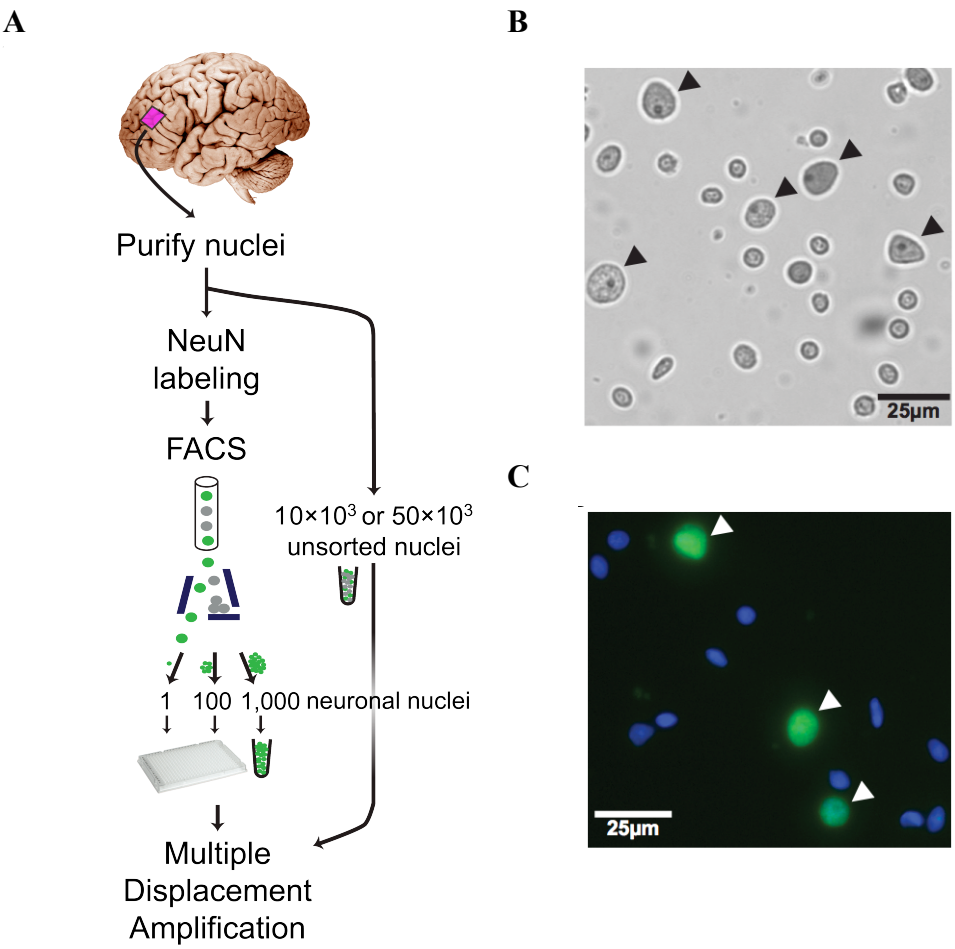


Figure 3-1. Isolation of neuronal nuclei from human brains.
(A) Schematic of single neuron isolation and whole-genome amplification.
(B) Purified nuclei from post-mortem human frontal cortex. Nuclei with neuronal nuclei morphology (large, prominent nucleolus) can be readily observed (arrowheads). Pyramidal shape in some large nuclei is reminiscent of pyramidal neuronal nuclei shape.
(C) NeuN (green) and Hoechst (blue) staining of cortical nuclei by fluorescent microscopy. White arrowheads indicate neuronal nuclei.

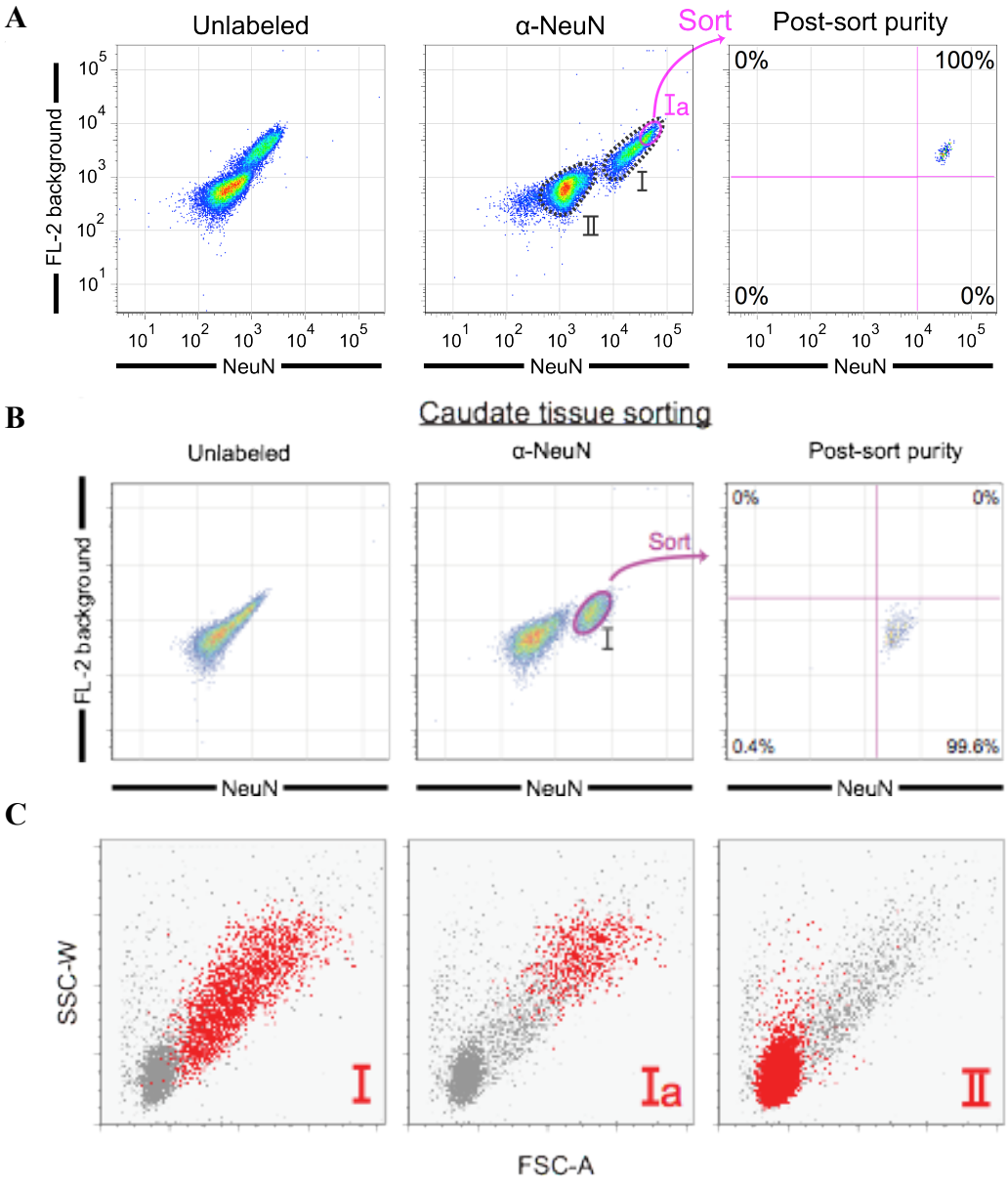


Figure 3-2. Fluorescence-activated cell sorting (FACS) of purified nuclei labeled with NeuN.
(A) FACS of cortical nuclei stained with NeuN shows two separable populations: $NeuN^+$ (population I) and $NeuN^-$ (population II). A subset of population I (Ia) consisting of large neuronal nuclei was sorted and reanalyzed, confirming sort purity. Two populations of nuclei are sometimes apparent without NeuN staining, due to the increased background staining of the larger population I nuclei. Fluorescence decrease of the sorted population on reanalysis is always observed due to photobleaching and washing of non-specific staining in the first sort.
(B) FACS of NeuN-labeled caudate nuclei. The entire $NeuN^+$ population (population I) was sorted and reanalyzed.
(C) Forward (FSC) and side (SSC) scatter backgating analysis of flow cytometry of human brain cortical nuclei. Red dots indicate events out of all events recorded (gray dots) from the specified population (I, Ia, and II). Population I nuclei have a distinct size distribution larger than population II, and population Ia nuclei are larger nuclei within population I.

(Gittins & Harrison 2004; Mills 2002), and indeed their nuclei often show a pyramidal shape (**Figures 3-2A**). We therefore sorted cortical nuclei within the top 25% NeuN/FL-2 fluorescence of population I (population Ia), which were the largest nuclei in population I (**Figures 3-2A**). We confirmed the neuronal and non-neuronal identities of the sorted populations by reverse transcriptase PCR (RT-PCR) and western blot analysis of additional neuronal (*SNAP25* and *SYT1*) and non-neuronal (*GFAP*, *AQP4*, and Olig2) markers (**Figures 3-3**). For every sort, a portion of the sorted nuclei was reanalyzed by FACS, confirming that nuclei remained intact during sorting and that sort purity was >98% (**Figures 3-2**).

Whole-genome amplification using MDA and GenomePlex WGA4

We first optimized MDA reaction condition for increased yield with varying concentrations of random hexamer, dNTP and phi29 polymerase. We found that dNTP is the most important limiting factor of reaction yield; and with sufficient dNTP present, increase of phi29 units improves the yield (**Figure 3-4A**). To balance between yield and the enzyme cost, we chose to follow the condition that produced the second-highest yield (15-20ug) at 50uM random hexamer, 2mM dNTP and 40U phi29 (**Figure 3-4A**). We then measured exogenous (non-human) DNA contamination in the reagents of the MDA reaction, finding negligible (< 1fg) exogenous DNA (**Figures 3-4B,C**) (Blainey & Quake 2011). Additional controls (see following section "Genome-wide coverage and amplification dropout rates of single neuronal genomes") excluded operator human DNA contamination. Quantitative MDA (qMDA) reactions further showed that, as the number of nuclei sorted in a well increased, the time-to-threshold-amplification decreased in a

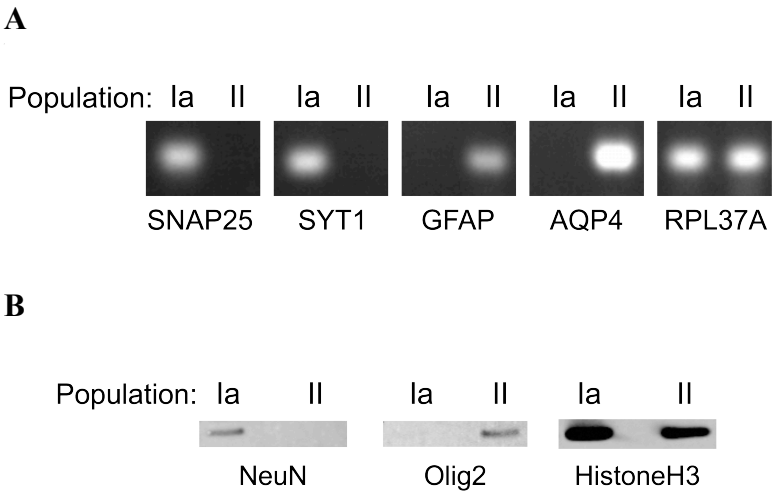


Figure 3-3. Post-FACS quality control on cell identity of sorted populations.
(A) RT-PCR confirming the neuronal and non-neuronal identities of populations Ia and II, respectively, by assaying for expression of nuclear RNA for two neuronal (*SNAP25* and *SYT1*), two astroglial (*GFAP* and *AQP4*), and input control (*RPL37A*) genes. RT-PCR and western blot experiments (Figures 1C and 1D) were performed with NeuN/Mef2c double labeling in which all $NeuN^+$ nuclei were $Mef2c^+$ (data not shown).
(B) Western blot analysis of NeuN and Olig2 (an oligodendrocyte marker), confirming neuronal and non-neuronal identity, respectively, of populations Ia and II.

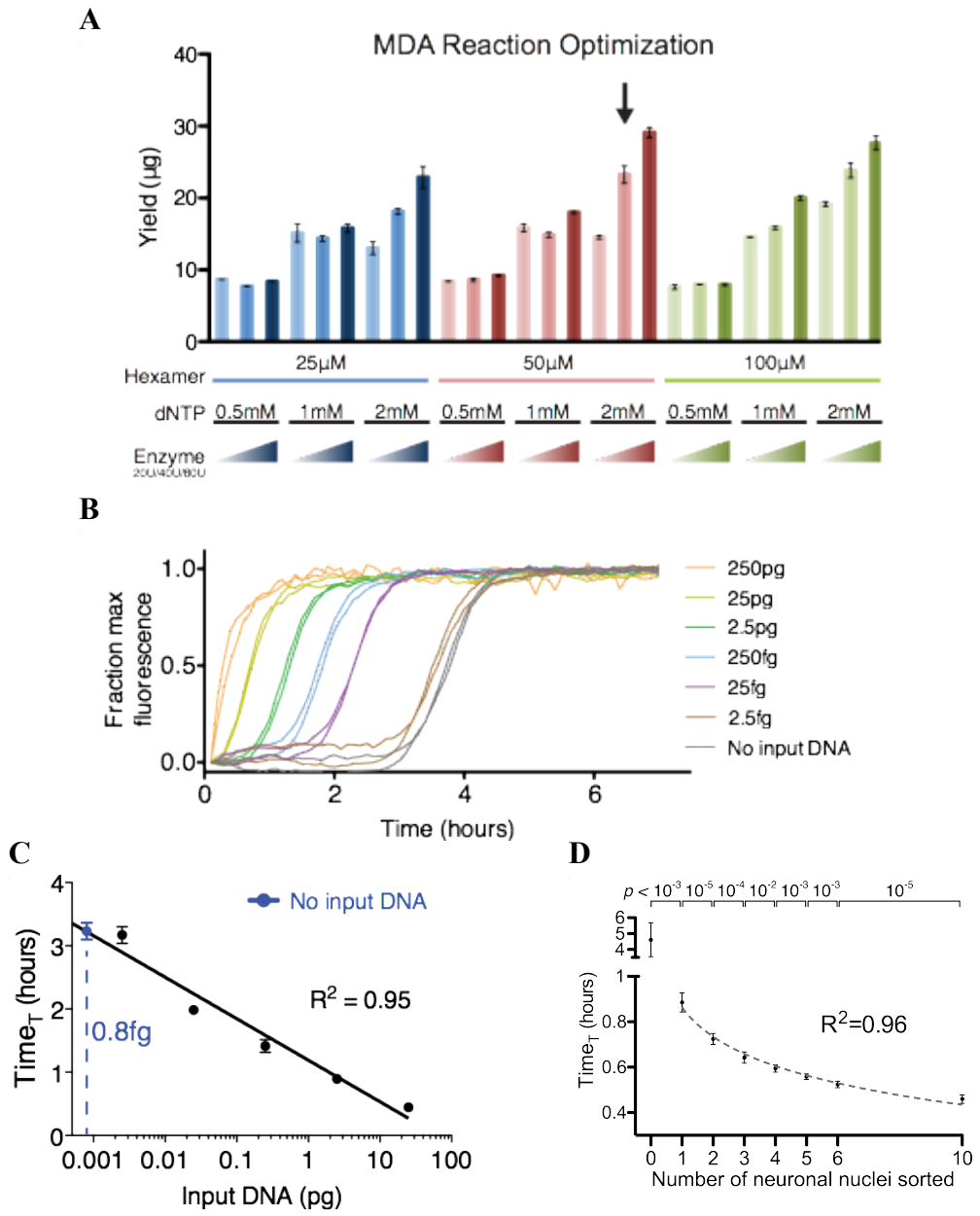


Figure 3-4. Optimization of single-cell MDA.
(A) MDA reaction yields of control human DNA amplified with varying hexamer, dNTP and phi29 polymerase concentrations (error bars ±1SD, n=2 per condition). Arrow indicates chosen reaction conditions for single neuron genome amplification.
(B) Real-time quantitative MDA monitoring of amplification reactions with varying lambda DNA input.
(C) A semi-log standard curve was fit to data from amplification curves in **Figure 3-4B** ($Time_T$, time to threshold amplification). No-input DNA reactions have ~0.8 femtograms of non-human exogenous DNA, negligible compared to 6.5 picograms of DNA in a single human nucleus (error bars ±1SD, n=2 per condition).
(D) Quantitative MDA reactions monitored in real-time confirm accurate sorting of the desired number of nuclei. The time to amplify to a threshold above background ($Time_T$, analogous to qPCR C_T value) is plotted on the y-axis (error bars ±1SD, n=7 or 8 reactions per condition). Points were fit to a semi-log line of slope -4.3, corresponding to 1.7-fold amplification per unit time.

step-wise manner (p <0.01 for each additional nucleus) (**Figure 3-4D**) (Zhang et al. 2006), confirming that the desired number of nuclei was correctly sorted in each well.

Because of concern that MDA would amplify the genome non-linearly, preventing analysis of aneuploidy and large segmental CNVs, we also explored the GenomePlex WGA4 method used by Navin *et al.* (2011) to profile copy numbers from single cancer cells. WGA4 starts with initial fragmentation and adaptor ligation of the single cell genome, followed by PCR amplification (**Figure 3-5A**). The fragmentation leads to non-lapping amplicions, and thereby PCR-based quality control assessments cannot be appiled to WGA4 amplified samples. Instead, per manufacturer's recommendation, we ran a small fraction of amplified samples on 1.5% agorase gel to examine the fragment size distribution and relative yield (**Figure 3-5B**) and quantified the final yield by nanodrop (data not shown) to confirm the successful amplification of each sample and the absence of external contamination of each preparation from the 0 cell negative control.

We concluded that our procedure can sort and amplify single neuronal genomes from human brains with high purity and in a high-throughput manner. For the analysis of SNPs, SVs and retrotransposon insertions, MDA amplified single cells were used. For the analysis of aneuploidy and segmental CNVs, both WGA4 and MDA amplified cells were used and compared (see detailed analysis and discussion in section "Amplification linearity of single cell genomes" in Chapter 4).

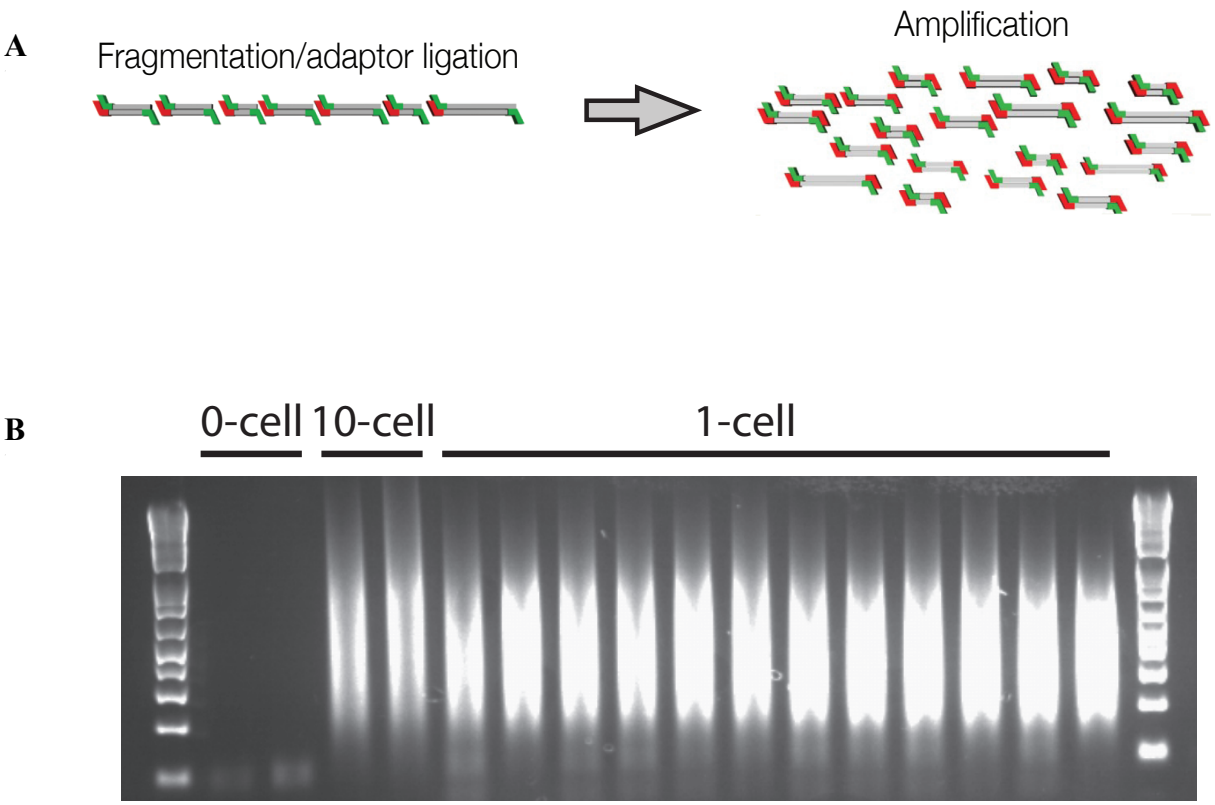


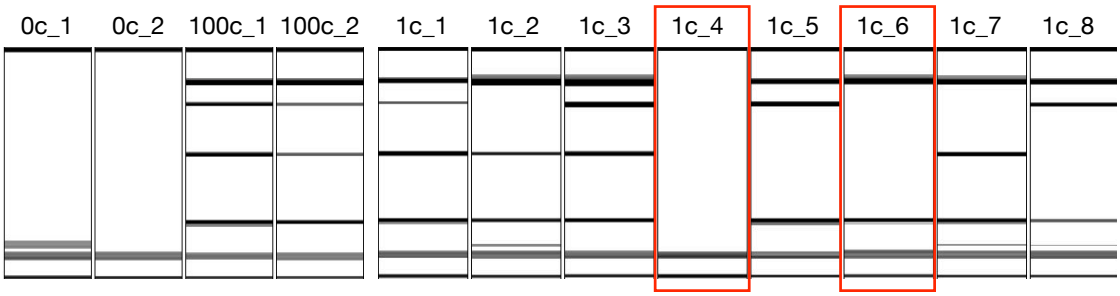
Figure 3-5. Quality assessment of GenomePlex WGA4 single cell whole-genome amplification.
(A) Schematic of the method.
(B) Agorose gel of amplified products from 0 cell, 1 cell and 10 cells, confirming succeussful amplification with fragments ranging between 150-1000bp, and free of external contamination based on 0-cell negative control.

Genome-wide coverage and amplification dropout rates of MDA single neuronal genomes

To be able to study a wide variety of mutational types from amplified single cell genome, it is important to determine what fraction of the genome has been consistently amplified by the method. We took multiple approaches, including multiplex PCR, low coverage sequencing, SNP-chip array, Identifiler fingerprinting and retrotransposon screening to estimate the allelic and locus dropout rates of single cell MDA at different scales ranging from single nucleotide from SNP-chip to ~500kb regions from low coverage sequencing. These results are largely consistent with each other and lead us to conclude that the single cell MDA method we developed recovers >90% of the single cell genome and thereby is suitable for studies of a wide spectrum of somatic mutational events (application on L1 retrotranspositon prolifing, aneuploidy analysis, and whole-genome sequencing for SNPs and SVs that will be discussed later in the chapter and in chapter 4).

We first evaluated the genome-wide coverage and reproducibility of our MDA single cell genome amplification. In an initial 4-locus multiplex PCR quality control, 97% of sorted single neurons showed amplification of at least 3 of the 4 loci, indicating that their genomes were successfully amplified and suitable for further experiments (**Figure 3-6A**). Low-coverage sequencing of MDA samples that failed the mutiplex PCR quality control confirmed that they were not suitable for downstream analysis because many genomic regions were amplified poorly, while a small fraction of the genome was overamplified up to 50-fold (**Figure 3-6B**). We then performed low-coverage whole-genome sequencing (**Figure 3-7A**) of eight randomly chosen single neurons (0.35x

A



B

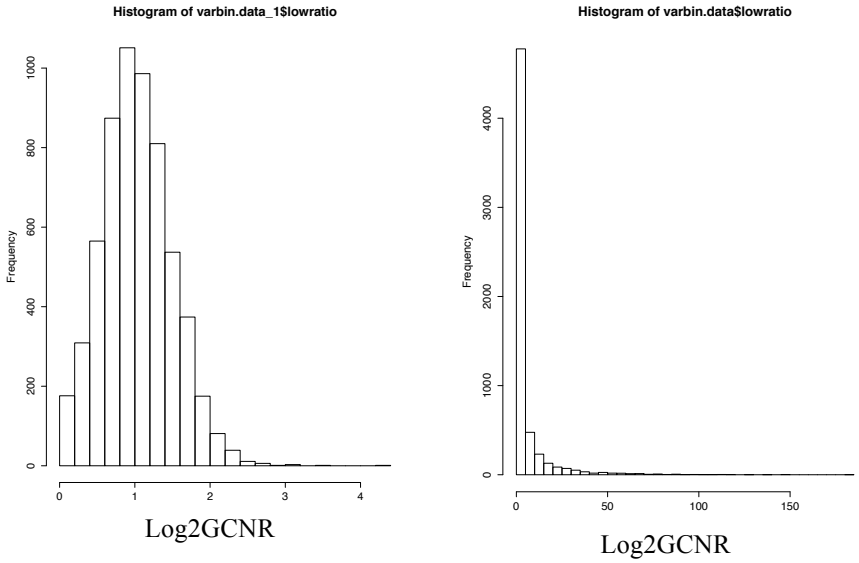


Figure 3-6. Multiplex PCR quality control on MDA amplified single cell samples.
(A) Multiplex PCR on 4 randomly selected loci from the human genome. 0 cell controls gave no amplified bands, confirming that the preparations are free of human DNA contamination. >80% of the sorted wells give at least 1 band, indicating a sorting success rate greater than 80%. Among the sorted wells, 97% give 3 or 4 bands, indicating a high success rate of MDA whole-genome amplification. Samples that yielded only 1 or 2 bands are regarded as poorly amplified and were excluded from further analysis. Red square marks wells with no cells sorted.
(B) Histograms of relative copy number ratio to euploid genome (normalized by read depth and GC content) of 6,000 equal-read bins across the whole genome. A representative sample passing the multiplex QC (left) shows a normal distribution with median = 1 (expected for euploid genome); a representative sample that failed the multiplex QC (right) shows a skewed distribution with most bins (~90%) showing copy number ratio close to 0 and the rest of the bins showing copy number ratio up to 50-fold.

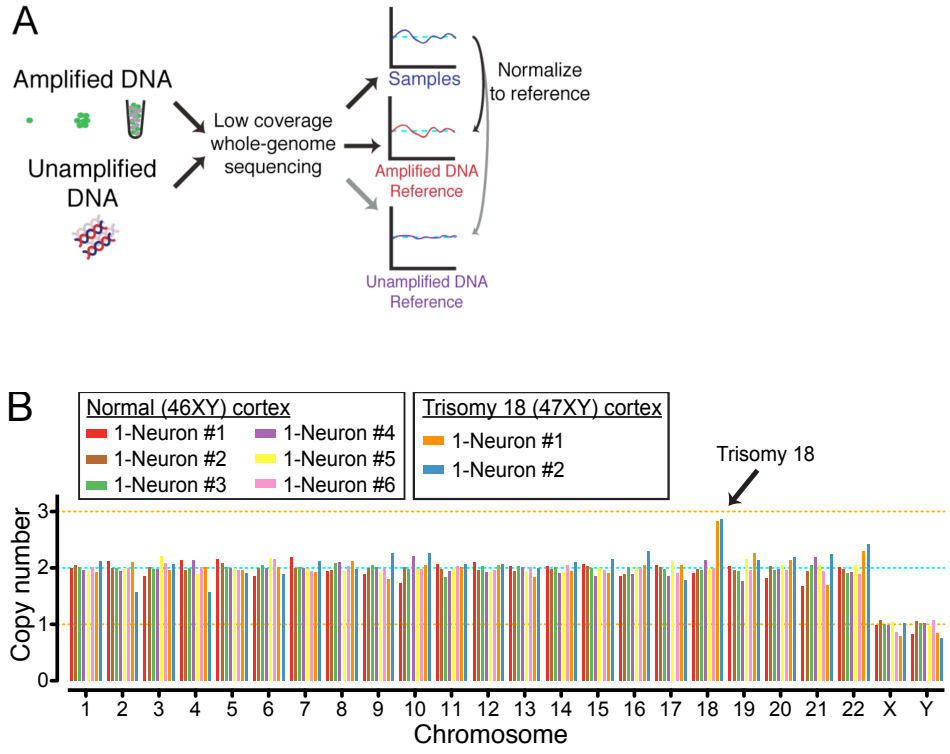


Figure 3-7. Whole-chromosome copy number analysis of MDA amplified single neurons.
(A) Schematic of the low-coverage whole-genome sequencing method.
(B) Chromosome copy numbers of single cortical neurons from normal (UMB1465, 46XY) and trisomy 18 (UMB866, 47XY,+18) individuals. Copy numbers are normalized to the median copy number of each chromosome across the 8 single neurons, with autosomes adjusted to a median copy number of 2. Orange lines denote ±1 copy.

average coverage), six from a normal male individual (46XY) and two from a male trisomy of chromosome 18, as well as unamplified and MDA-amplified bulk reference samples. The two neurons from the trisomy 18 individual showed the expected increase in chromosome 18 copy number, and the six single neurons from the normal individual were all euploid, confirming that intact nuclei can be sorted and that all chromosomes were amplified and that MDA is sensitive enough detect copy number changes at the chromosomal level despite previous concerns about its amplification linearity (**Figure 3-7B**). Counting sequencing reads across the genome in bins ~500kb in size (Navin et al. 2011; Baslan et al. 2012) revealed a systematic, regional amplification bias for all MDA samples, compared to unamplified bulk DNA, regardless of the number of nuclei amplified (**Figure 3-8A**). This regional bias in MDA amplification could be controlled for using any of the MDA samples as a reference (**Figure 3-8C**), indicating that most of the regional variability in amplification is inherent to MDA rather than reflecting the number of nuclei amplified. The rate of locus dropout (LD) rate at ~500kb resolution was estimated by counting the percentage of low-coverage sequencing bins with less than 1/16 of the copy number relative to an unamplified DNA reference, and was 2.0% for 1-neuron samples (**Figure 3-8B**). When using 100-neuron samples as a reference, LD in 1-neuron samples was lower at 0.05%, consistent with the finding that most regional amplification bias is inherent to MDA (**Figure 3-8B**). GC content partially accounts for the regional bias such that regions with high GC content are systematically under-amplified by MDA (**Figure 3-9A**). Regional bias resulting from GC content imbalance can be corrected by GC-normalization (see details in Methods) (**Figure 3-9B**).

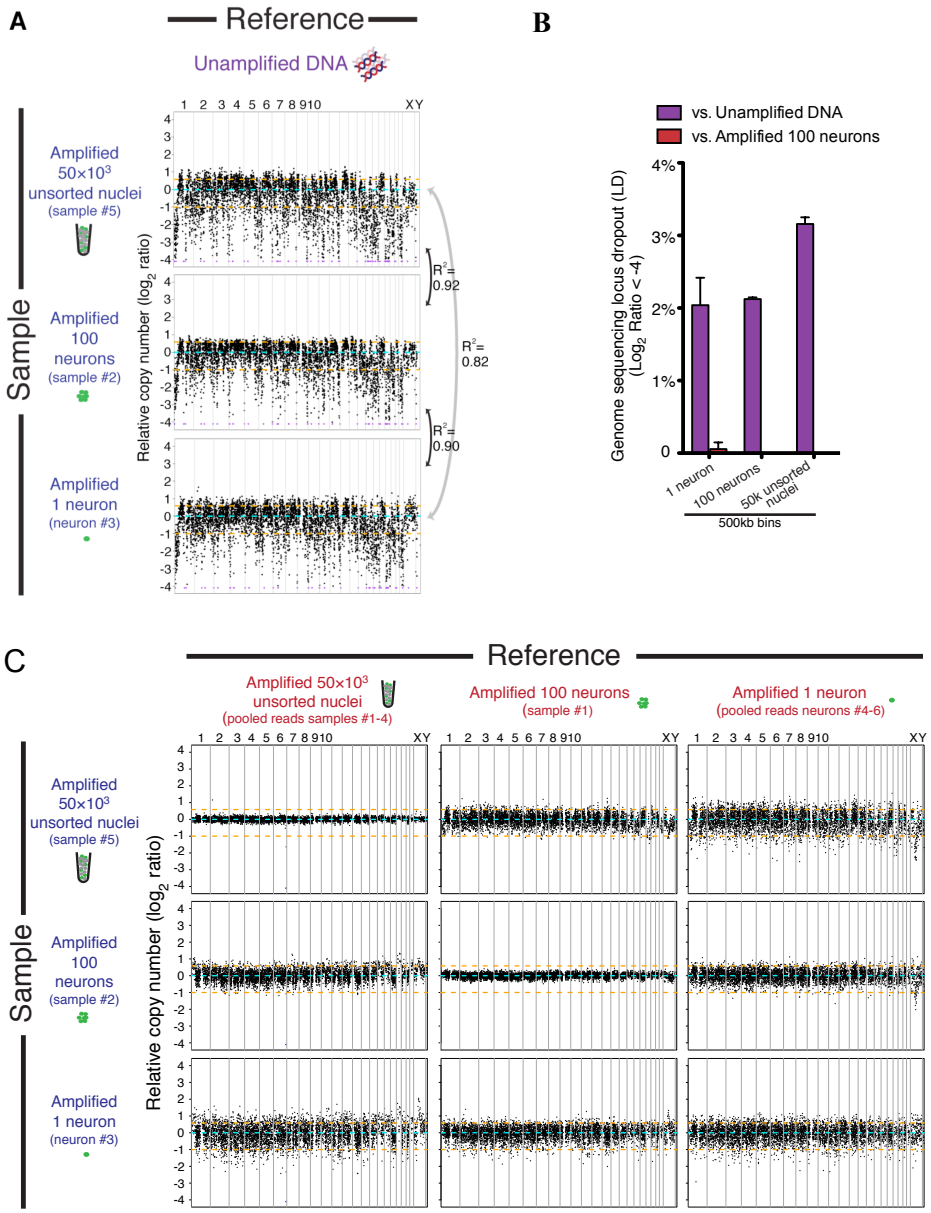


Figure 3-8. Low coverage whole-genome sequencing of MDA amplified single neuronal genome.
(A) Copy number profiling in 6,000 equal-read bins of ~500kb in size, relative to an unamplified DNA reference, shows that MDA bias is consistent and reproducible regardless of the number of nuclei amplified. Correlations (R^2) between 50×10^3-nuclei, 100- and 1-neuron samples are shown. Purple points represent off-scale bins.
(B) Genomic locus dropout (LD) estimates from low-coverage sequencing normalized to the indicated amplified and unamplified references (left panel) (error bars ±SD, n=4 for 50×10^3 nuclei, n=2 for 100-neuron, and n=6 for 1-neuron groups).
(C) Higher-resolution copy number profiling in 6,000 equal-read bins of ~500kb in size shows that MDA bias can be corrected by normalization to an MDA-amplified reference. Orange lines denote ±1 copy, and purple points indicate off-scale bins.

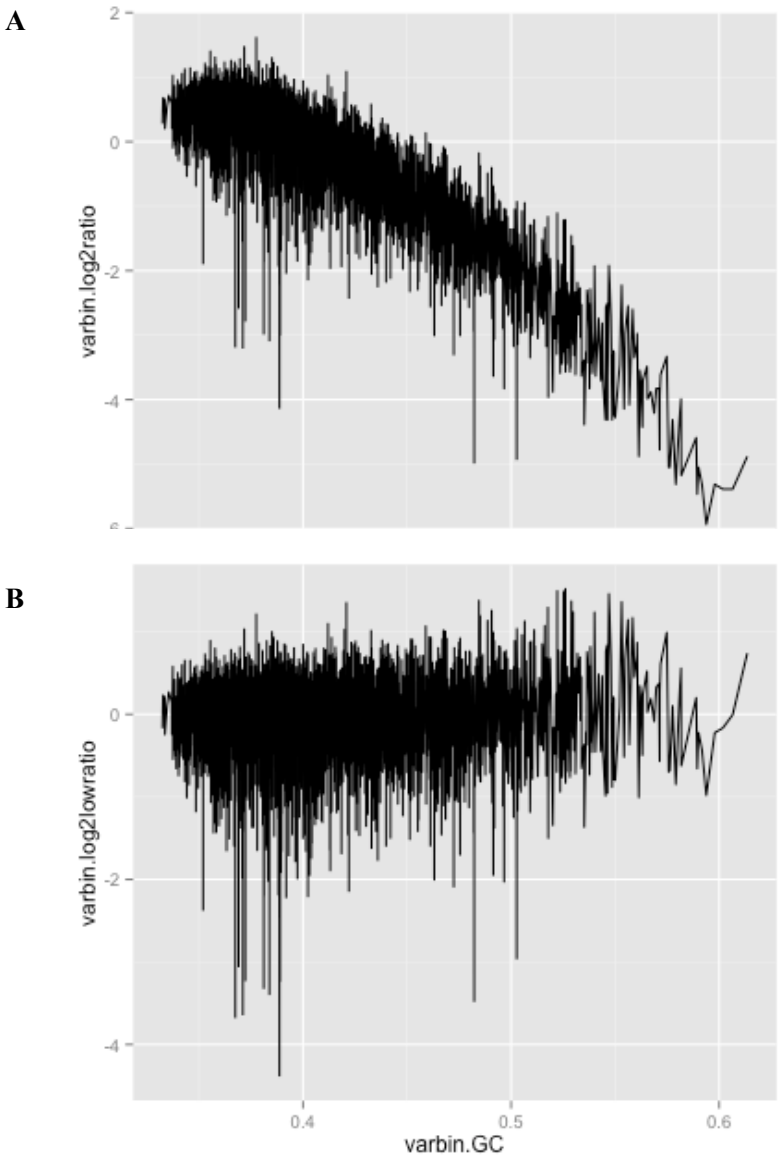


Figure 3-9. Normalization of MDA amplification bias by GC contents.
X-axis represents the GC content of the 6,000 equal-read bins across the genome; y-axis represents to the log2 ratio of relative copy number to euploid reference genome of each bin.
(A) Before GC-normalization, GC content negatively correlates to the log2 ratio suggesting that high GC regions are systematically under amplified by MDA.
(B) GC bias of the amplification can be corrected by GC-normalization (see details in Methods section).

A

Allelic dropout	Reference sample: 1465-cortex Unamplified Bulk DNA	Reference sample: 1465-lung Unamplified Bulk DNA
1465-cortex Unamplified Bulk DNA		0.00
1465-lung Unamplified Bulk DNA	0.00	
1465-cortex 1-neuron #6	0.08	0.08
1465-cortex 1-neuron #2	0.08	0.08
1465-cortex 1-neuron #3	0.09	0.09

B

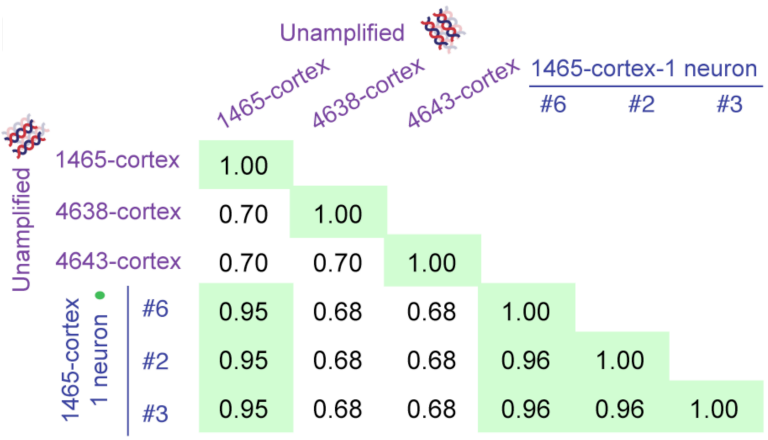


Figure 3-10. SNP-chip analysis on 3 MDA amplified single neurons.
(A) Allelic dropout (AD) rates of 3 single neurons against unamplified bulk DNA from the same individual at single base pair resolution.
(B) Fraction of genotypes by SNP microarray that are concordant between 3 single neurons and bulk DNA confirms the single neurons derive from the correct individual.

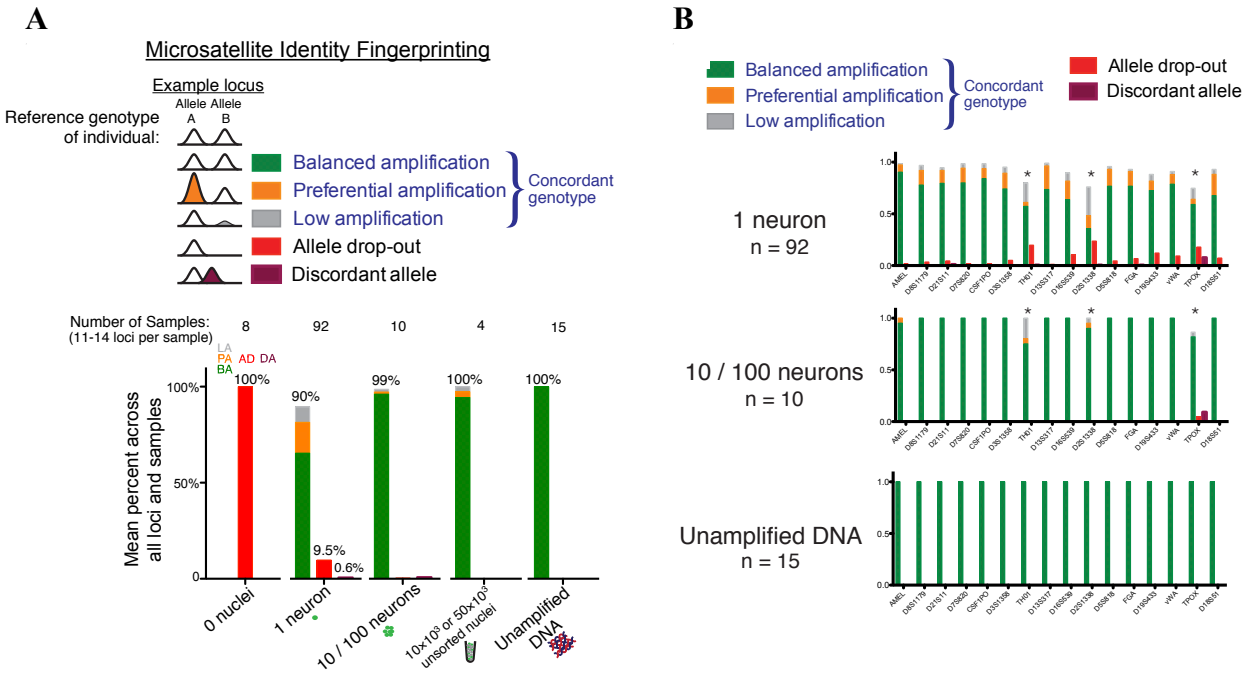


Figure 3-11. Identifiler fingerprinting analysis on bulk and MDA amplified samples.
(A) Identifiler fingerprinting confirms the single neurons derive from the correct individuals, and measures allele preferential amplification (PA), low amplification (LA), allele dropout (AD), and discordant allele (DA) rates.
(B) Per-locus Identifiler results for different sample types. Asterisks mark consistently under-performing loci, indicating that they reside in regions that do not amplify well by MDA.

In order to use single-neuron sequencing for somatic mutation detection, amplified genomes must reflect the diploid genotype (both alleles) of genomic loci. We therefore quantified the fraction of genomic loci that failed to amplify one (allelic dropout, AD) or both alleles (locus dropout, LD). Loss of one allele, AD, was measured by SNP microarray genotyping (**Figure 3-10B**) and with a panel of 16 highly polymorphic microsatellite markers (Identifiler fingerprinting) (**Figure 3-11A**). AD measured by SNP microarray (for >60,000 loci that are heterozygous in the bulk DNA and called with high confidence in both the reference and sample) was 8-9% in 3 single neurons (**Figure 3-10B**), and AD of 92 single neurons measured by Identifiler of 92 single neurons across 1,183 heterozygous loci, was 9.5% (**Figure 3-11A**), consistent with previous estimates (Hou et al. 2012). Some dropout tended to recur at specific loci even in MDA-amplified 100- and 1000-neuron samples (**Figure 3-11B**), reflecting the systematic bias of MDA for amplification of specific loci. Loss of both alleles, LD (locus dropout), was approximately 2.3% in the 92 single neurons assayed by Identifiler.

These low rates of AD (~10%) and LD (~2%) at various scales demonstrate comprehensive and reproducible amplification of single neuronal genomes by MDA, and suggest that genome-wide profiling of specific types of mutational events such as L1 insertions in single neurons could capture up to 90% of retrotransposon insertions per cell. These genotyping controls also excluded operator contamination, since all amplified single neuronal genomes tested were concordant with the bulk reference (**Figure 3-10B** and **3-11**).

Identification of somatic L1 retrotransposon insertion from single neurons

We carried out genome-wide L1 retrotransposition profiling on 300 MDA amplified single neurons from cortex and caudate of three neurologically normal individuals to identify somatic L1 insertions and to quantify the insertion rate (the estimation of somatic insertion rate is discussed in detail in Gilad Evrony's thesis). A method named L1Hs insertion profiling (L1IP) adapted from Ewing and Kazazian (2010), was used to profile L1Hs insertions genome-wide. L1Hs is the only retrotransposon subfamily that remains self-autonomously active in the human genome. There are 800-850 copies of L1Hs in each individual, of which ~600 copies are fixed in the human population (referred to later as known reference, KR) and 100-200 copies are population polymorphic (referred to later as known non-reference, KNR) (Ewing & Kazazian 2010; Iskow et al. 2010; Stewart et al. 2011; Hancks & Kazazian 2012).

We confirmed that we are able detect the expected absolute number of insertions: the mean number of KR, KNR and unknown insertions (UNK) per bulk DNA sample was 689, 113, and 43, respectively, compared to 628 KR and 152 KNR/UNK insertions found on average in a previous study (Ewing & Kazazian 2010). 605, 87 and 47 KR, KNR, and UNK insertions were found on average in the 300 single neurons profiled, reflecting a technical sensitivity >80% from the single neuron genomes. Furthermore, only 4 out 300 neurons profiled were poor quality outliers, demonstrating the high quality and consistency of the method (**Figure 3-12**).

In order to validate L1-IP predicted insertions, we optimized a 3' junction PCR validation method (3'PCR) (**Figure 3-13A**), and further used it to directly measure allelic dropout (AD) and locus dropout (LD) of L1Hs insertions in amplified single neurons. The technical sensitivity of the 3'PCR validation method (i.e. 3'PCR detection rate of

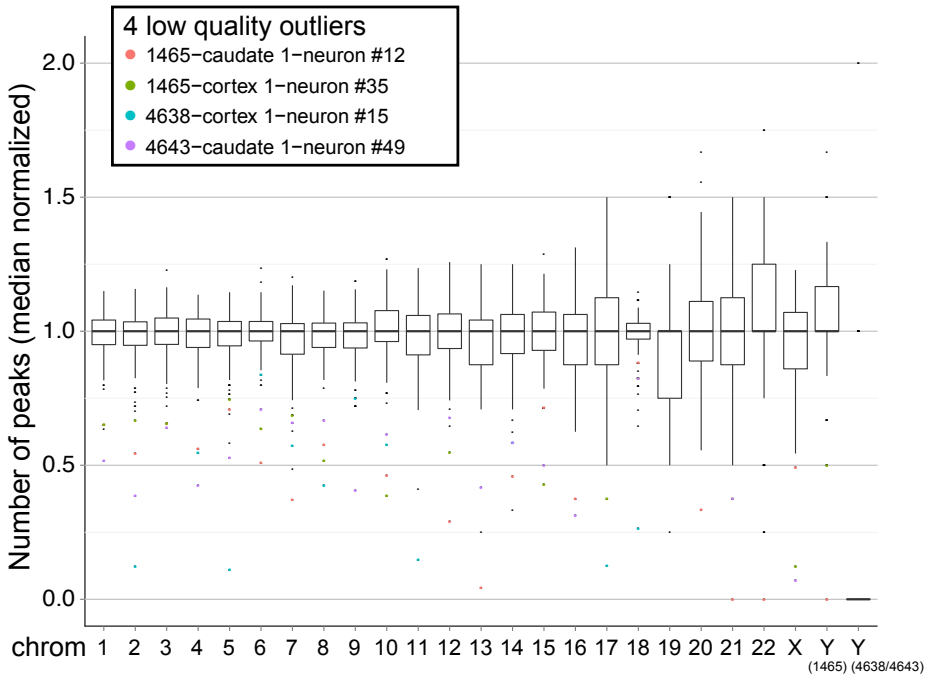


Figure 3-12. Box plot of number of L1IP insertion per chromosome, for all 1-neuron samples in the study.
Outliers are represented by black dots. 4 consistent outlier low-quality samples are colored.

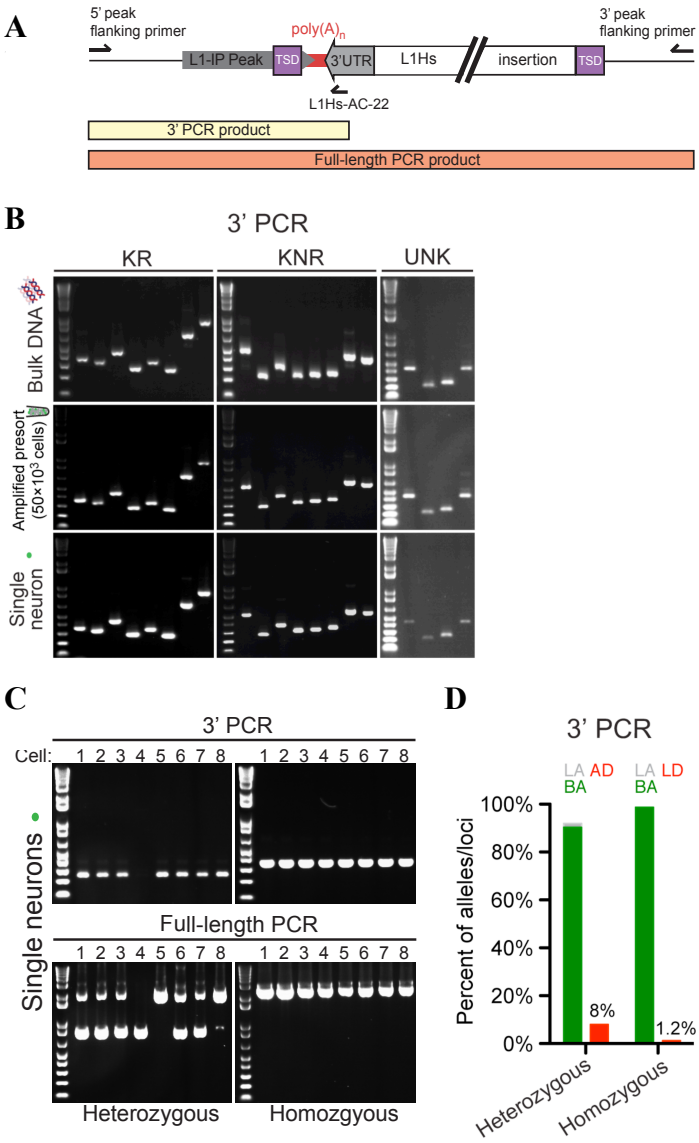


Figure 3-13. Estimation of AD and LD based on PCR validation of L1Hs insertions from single neurons.

(A) Schematic of the method.

(B) Representative gel images of 3'PCR of 20 different germline insertions (8 KR, 8 KNR, and 4 UNK).

(C) Representative gel images of 3'PCR and FL-PCR of 1 homozygous and 1 heterozygous L1Hs germline insertion in 8 different single neurons. The upper band in FL-PCR of the heterozygous insertion is the allele with the transposon insertion, and the lower band is the allele with no transposon insertion. Although the majority of cells have both alleles evenly amplified, AD of the insertion allele can be seen in some cells (e.g. neuron #4), which correlates with absence of 3'PCR product in the same cell. AD of the allele with no insertion can be seen in neuron #5. Neuron #8 had preferential amplification of the insertion allele.

(D) 3'PCR quantification of AD and LD in 1-neuron samples (n=83), of 3 heterozygous and 3 homozygous L1Hs insertions. AD and LD are quantified for heterozygous and homozygous insertions, respectively. BA, balanced amplification; LA, low amplification; AD, allelic-dropout; LD, locus-dropout.

true germline insertions) was important to determine first, in order to estimate at what rate true insertions found by L1-IP fail to validate by 3'PCR. This was assayed by 3'PCR of 64 known germline insertions (33 KR and 31 KNR) in unamplified bulk DNA, and in amplified, unsorted 50,000-nuclei and 1-neuron samples. In 1-neuron samples, 3'PCR detected 94% of known germline insertions with the first primer pair attempted, and the remainder were validated successfully with redesigned primers, This detection rate was not significantly different between amplified and unamplified samples (**Figures 3-13B**). 3'PCR can therefore sensitively detect L1Hs insertions in amplified single neuronal genomes. 3'PCR also successfully validated, in both bulk and 1-neuron samples, 12 out of 12 unknown (UNK) germline candidate insertions that we tested (**Figures 3-13B**), confirming that L1-IP can identify unknown insertions. AD of L1Hs insertions was then estimated by 3'PCR of 3 heterozygous insertions in a larger number of 83 single neurons (**Figures 3-13C**), finding 8.0% AD (20/249 alleles), consistent with previous estimates (**Figures 3-13D**). LD estimated by 3'PCR of 3 homozygous insertions in the same cells was 1.2% (3/249 alleles) (**Figures 3-13D**). We concluded that L1-IP's high sensitivity to detect germline insertions in single neurons, our robust 3'PCR validation method, and direct confirmation of <10% L1Hs allelic dropout, allows us to confidently search for somatic L1Hs insertions genome-wide in single neurons.

We successfully validated 5 somatic L1Hs insertion candidates by 3'PCR, and these 5 were studied further by attempting to clone their full-lengths, and screening for their presence by 3'PCR across all single neurons sorted from the individual in which they were found. We successfully cloned the full-length of one of the five somatic insertion candidates (**Figure 3-14A**). This insertion was detected in our L1-IP data in

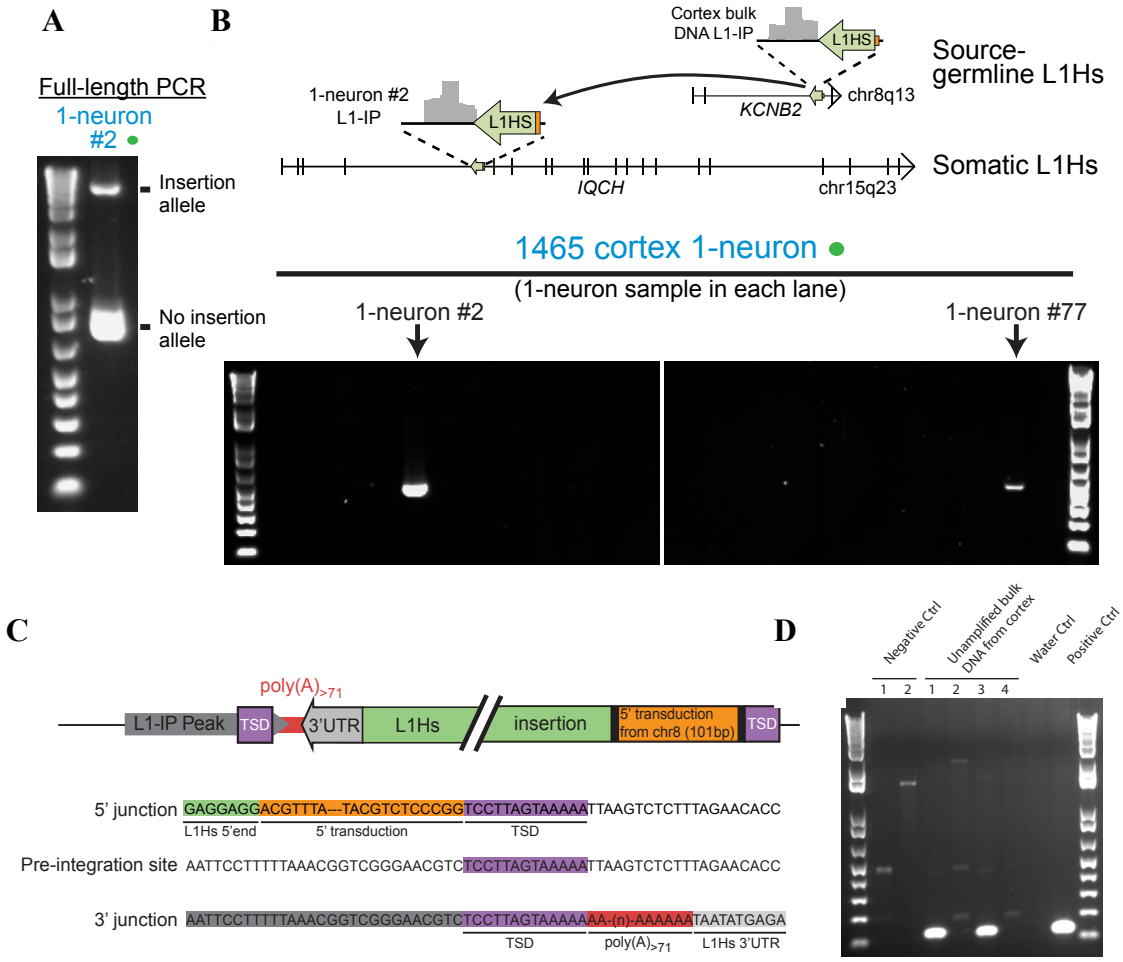


Figure 3-14. Identification of the first somatic L1 insertion in human brain.
(A) Gel image of full-length (long-range) PCR cloning of the validated L1Hs somatic insertion shown in Figures 6C-E (L1-IP peak ID chr15_67625710_plus_0_0), in sample 1465-cortex 1-neuron #2.
(B) Location of the somatic L1Hs insertion (L1-IP peak ID chr15_67625710_plus_0_0) in antisense orientation in intron 4 of the gene *IQCH*, and the corresponding L1-IP peak in 1465-cortex 1-neuron #2. The insertion's target site duplication coordinates are chr15: 67,625,702-67,625,714 (hg19). A 5' transduction (orange) identified the source L1Hs on chr8: 73,787,792-73,793,823. Representative gel images from a 3'PCR screen of 83 1-neuron samples from individual 1465 cortex (24 1-neuron samples shown). The two cortical 1-neuron samples (#2 and #77) found to have the insertion are shown. 1-neuron #77 was found to have the insertion only in the 3'PCR screen since it was not profiled by L1-IP. 3'PCR product sequencing and full-length cloning confirmed the insertion had identical 5' and 3' breakpoints and TSD in both neurons (#2 and #77).
(C) Structure of the L1Hs somatic insertion (chr15_67625710_plus_0_0), cloned by full-length PCR. Pre-integration TSD coordinates are chr15: 67,625,702-67,625,714 (hg19). 5' transduction (chr8: 73,793,824-73,794,027) from upstream of the source L1Hs (L1Hs-KR-chr8_73787792) exhibited transcriptional splicing removing 103bp (chr8: 73,793,831-73,793,934) from the source sequence. Sequences of 5' and 3' junctions of the L1Hs somatic insertion show. Poly-A tail length (at least 71bp long) was not possible to determine exactly due to difficulty sequencing through homopolymeric regions.
(D) Optimized 5' PCR confirms the presence of the somatic insertion in unamplified bulk DNA at low level.

intron 4 of the gene *IQCH* (IQ motif containing H, chromosome 15), in neuron #2 from the cortex of individual 1465, and is a full-length, intact 6.1kb L1Hs with all the hallmarks of a bona fide L1Hs insertion: a target site duplication (TSD) (13bp), a poly-A tail (~71bp), and a 5' transduction (101bp) allowing us to trace its source to a full-length, population-polymorphic KR L1Hs on chromosome 8 (**Figure 3-14B, C**). The full-length sequence of the somatic insertion precisely matched the sequence of the source L1Hs. The insertion was found in 2/83 (2.4%) cortical and 0/59 caudate single neurons tested (**Figure 3-14B**). The insertion was detected at low-levels in L1-IP data of some unsorted-50,000-nuclei samples, as expected for a low-level mosaic insertion. After further optimization, we were able to amplify the insertion from unamplified bulk samples with more specific primers against the 5' junction of the insertion, definitively showing that the somatic insertion identified was not an artifact due to MDA (**Figure 3-14D**). Further characterization of the percent of mosaicism and clonal dispersion pattern of the somatic insertion in different brain regions of the individual is described elsewhere (see Gilad Evrony's thesis). The remaining four candidates were each found by 3'PCR only in the single neuron in which they identified by L1-IP. Three of the four had poly-A tails by 3'PCR product sequencing (the fourth had an indeterminate poly-A tail since the breakpoint was within a genomic poly-A). Our results illustrate the ability of single-cell sequencing to identify somatic L1Hs insertions and highlight the potential of single-cell sequencing to identify very low-level mosaic mutations in human tissue.

Discussion

We have performed the first single cell, whole genome analysis to study somatic variants from normal human tissue. The method we developed allowed us to directly study human subjects without further experimental manipulations, such as tissue culture, cell transformation or iPS cell reprogramming. This ensured that we would capture genetic variants *in vivo*, free of common artifacts associated with *in vitro* manipulation. Our method is not limited to studies of the central nervous system, but widely applicable to any other organs, from which fresh-frozen post-mortem tissues are available. Most current human genetic studies heavily rely on DNA material from fresh blood; this approach inevitably misses somatic variants that are absent or present at only low levels in the hematopoietic system. Our method broadens the accessibility of tissue-specific somatic variants by making use of post-mortem tissues, which are generally available from public resource such as The NICHD Brain and Tissue Bank. We found that although it is difficult to isolate intact whole cells from frozen specimens, the nuclei integrity of these tissue specimens is largely preserved (with exceptions to be discussed in Chapter 4) and thereby is sufficient for genomic studies. However, for studies that require whole cells, such as single-cell mRNA sequencing, would still require fresh tissues.

We performed by far the most comprehensive and rigorous quality control assessments among all single cell genomic studies (Navin et al. 2011; Hou et al. 2012; B. Xu et al. 2011; Fan et al. 2011; Wang et al. 2012; Zong et al. 2012; Lu et al. 2012). Using quantitative MDA, we confirmed the exact number of cells sorted into a well for subsequent whole genome amplification and quantified the external contamination of

reagents. We also carried out all the procedures with extreme caution by UV treatment of all the reagents (except for enzymes, dNTP and oligos), consumables (e.g. PCR striptubes, plates, microtubes, pipette tips, etc.) and equipments (e.g. pipettes), and performed all the procedures within sterile laminar flow hoods. We included negative controls by sorting fluorescent beads within each sample preparation; and indeed, we never detected positive product from negative control wells by multiplex PCR, demonstrating that our single cell preparations are free of external human contamination. The individual cell identities were further confirmed by forensic fingerprinting and SNP-chip concordance on a subset of randomly selected samples. The initial multiplex screen on all amplified samples allowed us to immediately exclude poor quality samples from downstream analysis. This cost-effective screening method not only saved us money from sequencing bad samples, but also prevented potential data misinterpretation. Comprehensive analyses of allelic and locus dropout rates at various scales by different methods consistently pointed to an allelic dropout rate of ~10% and a locus dropout rate of no more than 2% for MDA amplified single cell samples. These quality control experiments led us to conclude that MDA single cell whole genome amplification is compatible with studies on a whole spectrum of somatic mutation types, except perhaps for small scale CNVs. This conclusion is further supported by our recent data from high-coverage (>30X) whole-genome sequencing on MDA amplified single neurons. We believe that a standard and through quality-control pipeline is essential for all future single cell genomic studies to allow for informative comparisons across studies and across methods.

In addition, we have optimized our protocol to be highly compatible with any high-throughput workflow such as robotic liquid handlers; therefore, systematic

assessments of large number of single cells is possible with our methods given the expected further reduction of sequencing cost.

Our validation of a somatic L1Hs insertion with all the hallmarks of a bona fide retrotransposition event, including a 5' transduction identifying its source, confirms that somatic L1Hs insertions are present in the normal human brain. The very low-level mosaicism of this insertion, and its detection only in cortical neurons, further suggests that it may have occurred during cortical development. Further characterization of this somatic insertion revealed that it is detectable in unamplified bulk DNA at an estimated percent of mosaicism between 0.1-0.2% only in a focal region of the cortex (data not shown), further confirming and highlighting the robustness of single cell sequencing in identifying low level mosaic somatic events.

Although we successfully demonstrated the possibility of using single-cell sequencing to detect and quantify low-level somatic variants that would be otherwise missed by analyzing bulk DNA, we do realize a number of limitations of the current methods. First of all, all amplification methods introduce errors including point mutations/small indels and chimeras. Although Phi29 is considered a "high-fidelity" polymerase with 3' → 5' exonuclease activity resulting in error rates around 10^{-6}- 10^{-5} (Wang et al. 2012), with an over million-fold amplification (6pg to ~15ug), a significant amount of point mutations are expected to be accumulated. Moreover, chimeras are known to be introduced during MDA amplification due to mechanisms such as branch migration and template switching (Lasken & Stockwell 2007). In our experience, the false positive rate of somatic variants from MDA amplified single cell samples is extremely high (i.e. 5/81 validated for somatic L1 insertion); therefore, secondary

validation steps that separate real somatic variants from technical artifacts are crucial for any single cell studies.

Materials and Methods

Tissue sources

Fresh-frozen post-mortem tissues of normal individuals and a trisomy 18 fetus were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland (Baltimore, MD). All tissues were frozen at -80°C with post-mortem intervals < 5 hours. Case UMB1465 was a 17 year-old male who died in a motor accident; case UMB4638 was a 15 year-old female who died in a motor accident; case UMB4643 was a 42 year-old female who died of cardiovascular disease; and case UMB866 was a 21 week-gestation fetus from an electively terminated pregnancy, with a 47XY,+18 (trisomy 18) karyotype. All tissue samples were confirmed as deriving from the correct individual with AmpFlSTR Identifiler Plus fingerprinting (Applied Biosystems).

Single-neuronal nuclei flow sorting and labeling

Nuclei were isolated and labeled for flow cytometry based on Spalding, et al. (2005) and Matevossian and Akbarian (2008). All procedures were performed at 4°C unless noted. 100-200mg of tissue was homogenized in a dounce homogenizer in lysis buffer (0.1% Triton X-100, 11% sucrose, 5mM $CaCl_2$, 3mM $MgAc_2$, 0.1mM EDTA, 10mM Tris-pH8, 50mM DTT) and ultra-centrifuged on top of a sucrose cushion (62% sucrose, 3mM $MgAc_2$, 10mM Tris-pH8, 50mM DTT) at 13krpm in an SW-28.1 rotor (Beckman Coulter). The pellet was resuspended in PBS+3mM $MgCl_2$ solution and

filtered through a 40μm cell strainer. Nuclear integrity, purity and concentration were assessed by light microscopy on a hemocytometer with light trypan blue staining.

For labeling with NeuN for flow cytometry, 1.2μg each of NeuN antibody (Millipore, MAB377) was pre-incubated with Alexa-488 goat α-mouse and Alexa-647 donkey α-rabbit antibodies (Life technologies) in 400μl PBS+3% BSA+3mM $MgCl_2$ solution at RT for 10 min. Initial experiments were performed with NeuN/Mef2c double-labeling, and since all $NeuN^+$ nuclei were also $Mef2c^+$ (data not shown) subsequent experiments were performed only with NeuN labeling. For NeuN/Mef2c double labeling experiments, 1.2μg each of NeuN and Mef2c antibody (Abcam, ab64644) was used. Alexa-647 secondary antibody was still used in addition to Alexa-488 in NeuN-only experiments to provide background signal (FL-2) relating to nuclear size. 200-500×10^3 nuclei were diluted in 1ml PBS+3mM $MgCl_2$ solution and incubated with the antibody mix at 4°C for at least 30 mins. Nuclei in Figure S1B were labeled with Alexa488-conjugated NeuN antibody (Millipore, MAB377).

Single nuclei were sorted at a maximum flow rate of 3.0 with a FACSAria II cell sorter at the Dana-Farber Hematologic Neoplasia Flow Cytometry core into 96-well (qMDA experiments) or 384-well (sequencing experiments) plates. Sorting into 384-well plates always left a gap of one empty well in all directions between single cells. 1,000 nuclei samples were sorted into microtubes. PBS+3mM $MgCl_2$ sheath fluid and sample chambers chilled to 4°C were used to help preserve nuclear integrity. SSC-H vs. SSC-W and FSC-H vs. FSC-W doublet discrimination gates and a stringent '0/32/16 single-cell' sort mask were used to ensure one and only one event was sorted per well. Initial experiments with DRAQ5 staining confirmed exclusion of doublets. Each sorted plate

contained negative (0 nuclei) and positive control (10 and 100 nuclei) wells. Amplified human DNA was never observed in negative control wells in quality control assays. Successful sorting of single nuclei into 384-well plates was monitored by quantitation of yield and with multiplex PCR (see below). Every plate's sorting success rate per well was >80%, and the sorting success rate per well across all plates was 94%.

RT-PCR and Western blots

5,000 nuclei from each population were sorted into microtubes. RNA was extracted using RNeasy (Qiagen) and cDNA synthesized with the Superscript III First-Strand Synthesis System (Life tech.), and in separate experiments protein lysates were used for western blots. RT-PCR primers were designed with Primer3 (Rozen & Skaletsky 2000) to span introns. Antibodies used for western blots were NeuN (Millipore, MAB377), Olig2 (Millipore, AB9601), and HistoneH3 (Abcam, ab1791) at 1:1000 dilution.

Single-neuron genome amplification by MDA

All work was carried out in a UV-treated laminar flow cabinet, and all surfaces, plastics and non-biologic buffers were UV-treated for at least 30 min. Reagents were added without touching the liquid surface to avoid losing parts of the single genome. Nuclei sorted into 384-well plates were sorted into 2.8µl lysis and denaturing buffer (200mM KOH, 5mM EDTA, 40mM DTT), and neutralized with 1.4µl neutralization buffer (400mM HCl, 600mM Tris-pH7.5). 15.8µl MDA reaction-mix was added to each well and incubated in a thermal cycler at 30°C for 16 hours (no lid heating), followed by 3 min at 65°C.

MDA reactions (Dean et al., 2002) were optimized for hexamer, dNTP, and phi29 polymerase concentrations by amplifying control human bulk DNA and assaying yield with Quant-iT Picogreen (Life tech.). All reaction conditions were confirmed to have high-molecular weight (>30kb) products by standard and alkaline gel electrophoresis (data not shown). Following optimization of MDA reaction conditions, MDA reagent concentrations used in reactions in this work were as follows: 1x RepliPHI phi29 reaction buffer (Epicentre), 50µM random hexamer 5′-dNdNdNdN*dN*dN-3′ (* = thiophosphate linkage) (IDT, Inc), 2mM each dNTP, 40U RepliPHI phi29 polymerase (Epicentre), and nuclease-free UV-treated water. Quantitative MDA reactions (Zhang et al., 2006) were monitored on a StepOnePlus real-time PCR instrument (Applied Biosystems) by addition of 0.1x SYBR Green I (Life tech.) and fluorescence was measured every 6 min for 7 hours. 0.5µl of MDA reaction products was diluted 1:50 for quality control assays (see below) and initial Picogreen quantitation. The remainder of the DNA was purified with AMPure XP beads (Beckman Coulter), treated with 10U mung-bean nuclease (NEB) at 30°C for 30 mins to debranch the MDA product structure (Zhang et al., 2006), cleaned-up with the DNeasy Blood & Tissue 96 kit (Qiagen) skipping the tissue digestion protocol steps, and assayed for final yield with Picogreen.

Single-neuron genome amplification by GenomePlex WGA4

WGA4 single cell whole genome amplification was performed according to the manufacturer's instructions (Sigma-Aldrich®), with slight modifications to the reaction volume and lysis conditions. All work was carried out in a UV-treated laminar flow cabinet, and all surfaces, plastics and non-biologic buffers were UV-treated at least 30 mins. Nuclei sorted into 96-well plates were sorted into either 5µl proteinase K lysis

buffer provided by the manufacturer's protocol or 4ul alkaline lysis and denaturing buffer (100mM KOH, 5mM EDTA, 40mM DTT), and neutralized with 1µl neutralization buffer (400mM HCl, 600mM Tris-pH7.5). Further fragmentation, adaptor ligation and PCR-amplification were performed according to the manufacturer's instruction with the reaction volume of all steps reduced by half.

Amplified products were quantified by nanodrop and the amplification quality as well as external contamination was assessed by 1.5% agarose gel electrophoresis of 5% of the total product.

MDA Amplified genome quality control

Dilutions (1:50) of MDA reaction products prior to cleanup were used for quality control assays. Every MDA reaction well, including negative and positive controls, was assayed for successful sorting of nuclei and to confirm absence of non-human and human DNA contamination with 2 methods: (a) Picogreen quantitation to measure yield and confirm success of controls (negative control reactions produce ~ ½ the yield of single nucleus reactions) and to estimate the percentage of wells that were successfully sorted; (b) multiplex PCR for 4 arbitrarily chosen loci from different chromosomes in the human genome was performed to exclude human DNA contamination in negative control reactions, to independently determine which wells contained a successfully sorted nucleus, and to exclude failed nuclei amplifying < 3 loci, likely indicating loss of significant parts of the genome during sorting or amplification. 96.8% of the wells into which a cell had successfully been deposited passed our 4-locus multiplex PCR quality control (3 or 4 of 4 loci amplified). Negative control wells and wells into which a nucleus failed to be sorted always produced both low yield by Picogreen and none of the

multiplex PCR bands. Multiplex primers were designed with PrimerStation (Yamada et al. 2006). Multiplex PCR reactions contained 5 μM of each primer, 1x HotStarTaq reaction buffer (Qiagen), supplemental 1.5mM $MgCl_2$, 0.2 μl HotStarTaq polymerase (Qiagen), 0.4mM dNTP, 2μl of 1:50 MDA reaction product, in 20μl reaction volumes. Thermal cycler conditions were: 94°C 15 mins, (94°C 1min, 68°C 1 min decreasing by 1°C every cycle, 72°C 1min, for 13 cycles), (92°C 1min, 55°C 1 min, 72°C 1 min, for 27 cycles), 72°C 10 min.

To further confirm the absence of any human DNA contamination and confirm the identity of sorted nuclei, additional quality control on a subset of 8-16 wells, including negative and positive controls, from each sorted plate, was performed using Identifiler muliplex genotyping of 16-microsattelite (STR) loci with the AmpFlSTR Identifiler Plus kit (Applied Biosystems) on a 3130xl Genetic Analyzer (Applied Biosystems). 1:50 dilution plates described above were further diluted to ~0.1ng/ul based on Picogreen quantitation for use in Identifiler asays. Unamplified bulk DNA genotypes were used as a reference. Loci homozoygous in an individual were excluded from preferential amplification (PA), low-amplification (LA), and allelic dropout (AD) calculations since they cannot be used to estimate per-allele PA, LA, and AD. 14 heterozygous loci were included in analysis for individuals 1465 and 4643 and 11 loci for 4638 (i.e. 11-14 heterozygous loci assayed in 92 single neurons = 1,183 loci assayed for the 1-neuron group). Genotypes of all samples were checked for concordance to the bulk reference genotype of the individual. Preferential amplification was defined as loci where the area under the trace of one allele was > 3x the area under the trace of the other allele. Low amplification was defined as callable alleles with traces of area < 1,000 fluorescence

units. Identifiler fingerprinting was performed on a subset of nuclei from every plate of nuclei sorted in this work.

Affymetrix SNP6 microarray genotyping (performed by Expression Analysis, Inc.) was performed on bulk DNA from cortex and lung tissue from individuals 1465, 4638 and 4643, and 3 single cortical neurons from individual 1465. Genotypes were called using the Affymetrix Genotyping Console with the Birdseed-v2 algorithm with standard settings. Genotypes called with confidence scores ≤ 0.01 in both the sample and reference were compared were used for analysis. Genotype concordance was calculated as: (# of loci with AA calls in both samples + # of loci with AB calls in both samples + # of loci with BB calls in both samples) / (total # of loci). The fractional allelic dropout (dropout rate) was calculated as: (# of heterozygous AB loci in the reference with AA or BB calls in the sample) / (2 × # of heterozygous AB loci in the reference). The fraction of discordant alleles was calculated as [(2 × # of loci with BB calls in the sample + # of loci with AB calls in the sample, where the reference is AA) + (2 × # of loci with AA calls in the sample + # of loci with AB calls in the sample, where the reference is BB)] / (2 × # of AA + 2 × # of BB loci in the reference). Depending upon the number of loci passing the confidence score threshold, for single-neuron versus bulk DNA comparisons between 250,000-350,000 loci were included for single-neuron versus bulk DNA comparisons in each comparison for genotype concordance, 60,000-75,000 loci for allelic dropout, and 200,000-300,000 loci for discordant allele calculations.

Whole-genome sequencing libraries

Whole-genome sequencing libraries for low-coverage sequencing were prepared from 1μg of DNA with the NEXTflex DNA sequencing kit (Bioo Scientific), and

barcoded for multiplexed sequencing at the Harvard Biopolymers Facility (Harvard Medical School) on a HiSeq 2000 sequencer (Illumina).

Sequencing copy number analysis

Raw reads were trimmed of low quality-score sequence and mapped to hg19 with Bowtie with -v 2 -m 1 --best --strata settings (Langmead et al. 2009). Chromosome copy numbers (**Figure 3-7B**) for each chromosome were calculated as the fraction of reads in each sample aligning to the chromosome, normalized to the median fraction of reads aligning to the chromosome across all 8 neurons. For autosomal chromosomes, these median-normalized relative chromosome copy numbers were multiplied by 2 to obtain absolute chromosome copy numbers, since both individuals were confirmed to have 46XY and 47XY,+18 karyotypes (i.e. 2 copies of each autosome except for chr18 in the trisomy 18 individual). The normal 46XY karyotype of the normal individual (UMB1465) was confirmed by comparison of 1465-bulk DNA low-coverage sequencing samples to in silico simulated reads from a 46XY genome (data not shown). The trisomy 18 individual (UMB866) was confirmed to have a 47XY,+18 karyotype from clinical karyotyping data at the NICHD Brain and Tissue Bank. Higher-resolution copy number normalization was performed by creating 6,000 (~500kb) bins spanning the entire genome with boundaries defined so that each bin contains an equal number of reads, B_{Ref}, in the reference sample (Navin et al., 2011). B_{Ref} was normalized to the reference sample's total read depth, T_{Ref}. The number of reads in the sample being analyzed in each bin, n, defined by the reference were then counted ($B_{Sample,n}$) and normalized to the sample's total read depth, T_{Sample}. For any bin 1…n, the relative copy number was calculated as $CN_n \frac{B_{Sample,n}/T_{Sample}}{B_{Ref}/T_{Ref}}$.

Replicate libraries from the 100-neuron #1 sample were highly reproducible ($R^2 > 0.9$ for

all pair-wise comparisons of copy-number profiles) and therefore pooled for analysis. Samples for analysis were always excluded from the pooled reference samples to avoid falsely lowering noise levels. The same was done for 100-neuron #2 replicate libraries. Copy number profiles were normalized to the global median of all bins with copy number >0.5 since dropout leads to an upward shift of the global midline (dropout bins lead to a proportional increase in reads in normal copy number bins in a given sample; global median normalization corrects this by shifting down all data points by an equal amount), and then $\log_2$ transformed relative copy numbers for analysis in R and visualization with scripts modified from the aCGH package (Fridlyand & Dimitrov, 2010).

For GC normalization, GC content of each genomic bin is calculated and plotted against the $\log_2$ transformed copy numbers ratio (**Figure 3-9A**). LOWESS smoothing was then applied to both the GC content and log2 ratio, followed by computing the normalization factor of each sample by logistic regression. GC-normalized log2 ratios show no bias to GC content of the respective genomic bins (**Figure 3-9B**). The algorithm was adapted from Baslan *et al.* (2012).

L1 insertion validation

Batch primer design

A custom primer design pipeline for L1Hs insertion validation was programmed in Excel, Galaxy, and Perl. L1-IP peak coordinates were used to define 750bp flanks 5' and 3' of the L1-IP peak, in order to design primers that flank the candidate insertion. The L1-IP peak 5' flank coordinates used to search for primers were 800 to 50bp upstream of the 3' end of the peak, and the 3' flank coordinates used were 400 to 1150bp

downstream of the peak. For peaks matching KR insertions, the 3' flank coordinates were the 750bp upstream of the KR insertion. These coordinates were used to extract genomic sequences both from an unmasked hg19 reference and an hg19 reference masked for non-unique 20bp sequences using the Duke Uniqueness track available in the UCSC genome browser. Both sets of sequences were used in parallel to search for high quality PCR primers in each flank in Primer3, with the following settings: target product size range: 301-850bp with preference for shorter amplicons, minimum primer size of 20bp, and a check for human repeat mispriming (remaining parameters are the default settings in the web version of Primer3). The batch primer design scripts then perform additional quality control on Primer3 primer results by checking the number of hits of each primer in the genome and the number of predicted PCR products using the Blat and in silico PCR functions of the UCSC genome browser. Primers with 1 genome hit and 1 predicted product were chosen from the non-uniqueness masked primer design results. Next, primers for peaks without primer pairs matching these criteria were chosen from the uniqueness-masked primer design results, again requiring at most 1 genomic match for the primers and predicted product. Primers for peaks without such primer pairs were then chosen allowing the L1-IP peak 3' flank primer to have 2 genomic hits. Primers for peaks still remaining without adequate primer pairs were manually designed with the aid of the Duke uniqueness track and Primer3. All primers were purchased from IDT.

3' junction PCR and full-length (FL) PCR validations

Two types of PCR were used for L1Hs insertion validation and characterization: 1) 3' junction PCR (3'PCR) with one primer specific to L1HS (L1Hs-AC-22) and the 5' peak flank primer (upstream to the peak), used to verify the presence of the predicted

insertion; and 2) long-range full-length PCR (FL-PCR) with the 5' and 3' peak flank primers to clone the entire length of candidate insertions and also to determine the zygosity (homozygous vs. heterozygous). All PCR products were run on 1% agarose gels and images were analyzed by ImageQuant TL software (GE Healthcare) to quantify the product sizes, relative intensities and absolute peak heights of the bands in an unbiased manner.

3'PCR: Positive 3'PCR reactions yield a single PCR amplicon within 150bp size of the predicted size. The predicted size was calculated as the distance from the 5' flank primer to the 3' end of the L1-IP peak plus 114bp, which is the distance of the AC primer location to the end of the L1Hs insertion, plus 50bp which is the approximate expected polyA tail length for recent polymorphic and disease-causing insertions (Beck et al. 2011; Hancks & Kazazian 2012). Peak coordinates do not necessarily precisely border the insertion in situations where seed sequences are not present adjacent to the insertion or low mappability prevents read mapping adjacent to the insertion, which leads to only approximate predicted sizes. The difference between observed to predicted band sizes was -1 ± 47 bp (SD) for 71 out of 76 insertions that validated in a PCR sensitivity screen (see below), supporting our ability to predict amplicon size within 150bp. Negative reactions yield no PCR product or in rare cases a band outside of the predicted size range.

FL-PCR: For heterozygous insertions genotyped by FL-PCR, two products are expected: a smaller product within 150bp of the predicted size without an L1Hs insertion, and a product ranging up to 6kb larger than the smaller product, depending on the size of the L1Hs insertion. Some candidates validated by 3'PCR may still yield only one band in FL-PCR at the expected no-insertion size range, since even long-range optimized FL-

PCR reactions can be biased toward smaller products. In this case, we conclude that the candidate insertion is heterozygous based on the presence of an amplicon corresponding to the absence of an insertion and 3'PCR validation of the presence of an insertion.

Previously published PCR protocols for L1Hs validation (Ewing and Kazazian, 2010; Iskow et al., 2010; Stewart et al., 2011) were adapted and optimized to maximize sensitivity and specificity for both unamplified bulk and MDA-amplified single-cell DNA:

Table 3-1. PCR protocols of 3'PCR and FL-PCR.

3'PCR master mix	**Amount**	**3'PCR program**			
5X GoTaq flexi buffer	4ul	Steps	Temperature	Duration	Cycles
MgCl2 (25uM)	1.2ul	1	95	5min	
dNTP (10uM)	0.4ul	2	95	30sec	
GoTaq Hot Start polymerase (Promega)	0.2ul	3	60	30sec	
AC-22 primer	0.8uM	4	72	1min	
Primer 5'	0.8uM	5			35X
DNA template	5ng	6	72	5min	
Total reaction volume	20ul	7	4	hold	
FL-PCR master mix	**Amount**	**FL-PCR program**			
10X LA Taq buffer	2ul	Steps	Temperature	Duration	Cycles
dNTP	3.2ul	1	94	90sec	
LA Taq (Takara Bio)	0.2ul	2	94	20sec	
Primer 5'	0.5uM	3	61	20sec	
Primer 3'	0.5uM	4	68	8:30min	
DNA template	10ng	5			32X
Total reaction volume	20ul	6	68	10min	
		7	4	hold	

PCR sensitivity and specificity calculations

The sensitivity and specificity of the 3'PCR validation method was assessed by performing validation on 64 high-confidence known germline insertions found by L1-IP in bulk tissues (33 KR and 31 KNR). In a separate experiment, 3'PCR validation was performed on 12 high-confidence unknown (UNK) germline candidate insertions found

by L1-IP in bulk tissues. Among these 76 insertions (64 known and 12 unknown), 31 were present in all 3 individuals in this study and 45 were absent from at least 1 of the 3 individuals (polymorphic).

Sensitivity: The sensitivity of the 3'PCR was 92% (70/76) for unamplified bulk DNA, 92% (70/76) for MDA-amplified unsorted (50 x 10^3 cells), and 93% (71/76) for MDA-amplified single-cell (one sample of each type assayed), demonstrating consistent sensitivity of 3'PCR for both unamplified and MDA-amplified DNA (**Figure 3-13B**). 2/6 of the insertions that failed validation did not have any visible PCR product, and 3/6 had product of the wrong size. The PCR sensitivities for the KR, KNR, and UNK insertions were 91% (30/33), 94% (29/31), and 92% (11/12), respectively, in both unamplified bulk DNA and unsorted-50k-nuclei amplified DNA, and 91% (30/33), 97% (30/31), and 92% (11/12), respectively, in 1-neuron amplified DNA. For candidates that failed the initial 3'PCR validation, up to 3 additional 5' peak flank primers were tested to differentiate PCR failure from false insertion predictions of the L1-IP pipeline. The PCR sensitivity after an additional second set of primers increased to 97%, 97% and 96% in bulk DNA, unsorted nuclei and 1-cell, respectively. After testing failed candidates on up to four sets of primers, the final sensitivities were 100% for bulk DNA, unsorted-nuclei and 1-cell samples, confirming that the initial loss of sensitivity was due to faulty primers.

Specificity: The specificity of the 3'PCR method for loci without predicted insertions was also determined by assaying for polymorphic L1Hs loci in individuals predicted not to have an insertion by L1-IP. These loci should not yield a band of the predicted size. The 45 polymorphic germline insertions were assayed in the individuals predicted by L1-IP to not have the insertions. The experiment was performed on one

sample each of bulk DNA, amplified unsorted-nuclei and single cell, from each of the individuals without the insertion. No false positive validations were observed in bulk DNA, unamplified-nuclei, and 1-cell samples (63 reactions each). 187/189 of the reactions had no band. 2/189 had a band >1kb larger than the predicted size in bulk and 1-cell samples of one individual, and failed validation for this reason. The specificity of the 3'PCR was therefore 100%.

L1Hs 3'PCR and FL-PCR single cell allelic and locus dropout rates

3'PCR sensitivity for one single-cell in the above experiment was the same as bulk and unsorted-nuclei DNA. However, the above experiment was based on only one single-cell and 46% of assayed insertions whose zygosity could be determined by FL-PCR were homozygous, such that this overall rate does not reflect the true AD and LD rates. Therefore, a more comprehensive assessment of AD and LD was carried out by 3'PCR genotyping of 3 homozygous (for LD estimation) and 3 heterozygous (for AD estimation) insertions in each of 83 single-neurons from individual 1465. Low amplification in 3'PCR was defined as callable alleles with peak height < 10,000 arbitrary fluorescence units. Peaks with height < 5,000 units were considered as allelic dropout as these are barely above background noise.

TOPO-TA cloning and Sanger sequencing

PCR products were sequenced either by direct Sanger sequencing of PCR products, or by TOPO-TA cloning (Life Tech.) of PCR fragments for subsequent Sanger sequencing. All Sanger sequencing was performed by Genewiz. Sequence traces were analyzed and assembled by Geneious.

References

Baslan, T. et al., 2012. Genome-wide copy number analysis of single cells. 7(6), pp.1024–1041.

Blainey, P.C. & Quake, S.R., 2011. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic acids research*, 39(4), p.e19.

Dean, F.B. et al., 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8), pp.5261–5266.

Evrony, G.D. et al., 2012. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell*, 151(3), pp.483–496.

Ewing, A.D. & Kazazian, H.H., 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome research*, 20(9), pp.1262–1270.

Fan, H.C. et al., 2011. Whole-genome molecular haplotyping of single cells. *Nature biotechnology*, 29(1), pp.51–57.

Gittins, R. & Harrison, P.J., 2004. Neuronal density, size and shape in the human anterior cingulate cortex: a comparison of Nissl and NeuN staining. *Brain research bulletin*, 63(2), pp.155–160.

Hancks, D.C. & Kazazian, H.H., 2012. Active human retrotransposons: variation and disease. *Current opinion in genetics & development*, 22(3), pp.191–203.

Hou, Y. et al., 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 148(5), pp.873–885.

Iskow, R.C. et al., 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7), pp.1253–1261.

Magavi, S. et al., 2012. Coincident Generation of Pyramidal Neurons and Protoplasmic Astrocytes in Neocortical Columns. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(14), pp.4762–4772.

Marcy, Y. et al., 2007. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS genetics*, 3(9), pp.1702–1708.

Matevossian, A. & Akbarian, S., 2008. Neuronal nuclei isolation from human postmortem brain tissue. *Journal of visualized experiments : JoVE*, (20).

Mills, S.E., 2002. Histology for Pathologists. *Health Perspect*.

Mullen, R.J., Buck, C.R. & Smith, A.M., 1992. NeuN, a neuronal specific nuclear protein in vertebrates. *Development (Cambridge, England)*, 116(1), pp.201–211.

Navin, N. & Hicks, J., 2011. Future medical applications of single-cell sequencing in cancer. *Genome medicine*, 3(5), p.31.

Navin, N. et al., 2011. Tumour evolution inferred by single-cell sequencing. *Nature Genetics*, 472(7341), pp.90–94.

Parent, A. & Carpenter, M.B., 1995. Human neuroanatomy.

Rehen, S.K. et al., 2005. Constitutional aneuploidy in the normal human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(9), pp.2176–2180.

Spalding, K.L. et al., 2005. Retrospective birth dating of cells in humans. *Cell*, 122(1), pp.133–143.

Stewart, C. et al., 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics*, 7(8), p.e1002236.

Wolf, H.K. et al., 1996. NeuN: a useful neuronal marker for diagnostic histopathology. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 44(10), pp.1167–1171.

Xu, X. et al., 2012. Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell*, 148(5), pp.886–895.

Zare, R.N. & Kim, S., 2010. Microfluidic platforms for single-cell analysis. *Annual review of biomedical engineering*.

Zhang, K. et al., 2006. Sequencing genomes from single cells by polymerase cloning. *Nature biotechnology*, 24(6), pp.680–686.

Chapter 4: Chromosomal Copy Number Analysis of Single Neurons from Normal and Hemimegalencephalic Brains

This chapter contains partially unpublished work as well as data presented in the manuscript "Somatic Activation of AKT3 Causes Hemispheric Developmental Brain Malformations", published in *Neuron*, April 12, 2012; 74: 41-48, and the manuscript "Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain", published in *Cell*, October 26, 2012; 151(3):483-496. The text and figures were modified to fit the format of this dissertation. Xuyu Cai led all the analysis on single cell copy number with assistance from Gilad Evrony, Princess Elhosary, Ben Hills and Bhaven Mehta. Ann Poduri performed the initial copy number screening on HMG brains. Gilad Evrony and Ben Hills performed the qPCR experiments on HMG-1 brain.

Summary

This chapter summaries our efforts at copy number analysis of single-cell amplified genomes, both at the chromosomal level and down to 2Mb segmental CNVs. A quality control metric was developed to assess sample quality and amplification linearity of single-cell whole-genome amplified samples. Two currently available methods for single-cell whole-genome amplification—including multiple displacement amplification (MDA) and GenomePlex Whole-Genome Amplification 4 (WGA4)—are compared at various aspects for their applications to copy number analysis. Single cortical neurons from 3 normal individuals, 1 trisomy 18 fetus and 1 hemimegalencephaly brain (HMG-1) are analyzed for somatic aneuploidy and segmental CNVs. We find that the majority of single cortical neurons analyzed, either by MDA or by WGA4, are euploid at the level of entire chromosomes, and even chromosome arms. We use single cell analysis to accurately determine the precise chromosome anomaly associated with HMG, defining tetrasomy of chromosome 1q in a patient with HMG, rather than the trisomy that had been predicted based on analysis of bulk tissue. We also find that few (<5%) cortical neurons from normal brain are grossly aneuploid at the level of chromosomes.

Introduction

In the previous chapter we have shown that single-cell whole-genome amplification by MDA successfully recovers ~90% of the genome, making it potentially amendable for studies of multiple types of somatic mutations, such as single-nucleotide variations (SNV), microsatellite repeats up to a few hundred basepairs, as well as L1 retrotransposons up to 6kb. Additional study has also demonstrated the successful use of single-cell MDA to study meiotic recombination in human sperm (Wang et al. 2012). However, copy number analysis on MDA amplified single

cells is known to be challenging given the inherent amplification noise and false positive dropouts introduced by MDA. Several array-based copy number studies of single cells amplified by MDA demonstrated the feasibility of applying single-cell MDA to copy number analysis, but also revealed significant challenges with signal normalization and false positive differentiation of these amplified samples (Vanneste et al. 2009; Cheng et al. 2011). We propose that next-generation sequencing-based approaches will improve the performance of copy number analysis on MDA amplified single cells as it provides a wider dynamic range for signal detection, as well as more flexibility for data manipulation. In addition, it provides a more cost-effective way to analyze a large number of samples simultaneously.

Technical Challenges

Copy number analysis relies on the change of signal intensity from the genomic regions with altered copy numbers. Such signal intensities are measured by the hybridization signal at each probe through array-based methods and by the number of read counts of each specified genomic segment (named genomic "bins" or "windows") through sequencing-based methods. Regardless of the methods, the key for successful single-cell copy number analysis is to have the amplified genome faithfully reflect the initial copy number states of the single-cell genome. This is defined as the amplification linearity or amplification uniformity. Non-linear amplification often occurs when amplifying from small copies of initial material, as any selective amplification in the initial rounds will become prominent copy number fluctuations in the end-product independent from the initial copy number states. Non-linear amplification leads to several technical artifacts, including 1) systematic amplification bias, 2) stochastic amplification noise and 3) false-positive dropouts. Systematic amplification biases give rise to regions that are consistently over or under-amplified specific to the amplification methods. One prominent factor

affecting systematic biases is the GC content, which can be corrected for by building-in GC-normalization into the analysis pipeline. Additional systematic biases include secondary and tertiary structures of the genome, repetitive regions and perhaps nuclear compartmentation. Stochastic amplification noise is caused by regional over- or under-amplification, resulting in random fluctuation in copy numbers of each genomic bin around its true CNV state. Since stochastic amplification noise is expected to happen randomly, no normalization can be applied. However, assuming its randomness, the average of several consecutive data points should reflect the true CNV state of the region, as long as the size of a true CNV is significantly larger than the size of genomic region represented by each bin. A common computational tool to correct for stochastic noise is copy number segmentation, which computationally determines genomic regions with altered copy number states supported by multiple bins (Olshen et al. 2004). One advantage of sequencing-based methods over array-based methods is that not only the number of data points supporting each CNV can be changed; the number of genomic bins determined by the size of each bin is also able to be manipulated. Stochastic noise can be caused by amplification itself, as well as subsequent sequencing analysis and data manipulation; therefore, it is important to keep track of the two variables by comparing amplified and unamplified samples in parallel with the same downstream analyses. It is conceivable that the amplification noise directly correlates with the sensitivity, specificity and resolution of copy number calls, and thereby needs to be closely monitored throughout the analyses. Lastly, false dropouts can happen randomly throughout the amplified genome at various sizes. In general, single copy dropouts are more common than two-copy dropouts, and smaller sized dropouts are more common than larger sized dropouts. False dropouts are particularly harmful for single-cell copy number analysis as they are mixed with true deletions and are the major source of false positive errors. Therefore, assuming

each of the single cells analyzed contains a comparable number of true somatic deletions, the dropout statistics can be monitored and used to exclude poor quality samples flagged by a significantly increased number of dropouts. Taken together, the three above-mentioned factors, systematic amplification bias, stochastic amplification noise and false positive dropouts, need to be monitored and compared throughout the copy number analyses for the assessment of amplification methods and sample qualities.

Alternative Amplification Methods for Single-Cell Copy Number Analysis

An alternative method to MDA for single-cell whole-genome amplification is a PCR-based method called GenomePlex WGA4 licensed by Sigma (See details in Chapter 3 *Single cell whole-genome amplification*). Several studies have successfully combined this method with next-generation sequencing to study CNVs in cancer single cells, as well as single cells isolated from *in vitro* human blastomeres at preimplantation genetic diagnosis (PGD) (Navin et al. 2011; Yin et al. 2013; C. Zhang et al. 2013). Although there has been no direct comparison prior to our study, WGA4 has been known to be the better method for copy number analysis with resolution potentially down to ~500kb (Navin et al. 2011; Baslan et al. 2012); however, its performance on non-cancerous postmortem tissue has yet to be addressed. The major drawback of the WGA4 method is its low genome coverage (<10%), which prevents the analysis of other mutational types from the same amplified samples. This is the major motivation for us to explore the possibility of using MDA for large-scale copy number analysis, such as aneuploidy. Additionally, a newer method named multiple annealing and looping-based amplification cycles (MALBAC) was developed recently (Zong et al. 2012), after the completion of our current study. The MALBAC method claims to have superior performance on copy number analysis over MDA, as well as better genome coverage than WGA4. Therefore, it could potentially be used as an all-in-

one method for single-cell somatic mutation detections. However, based on our data reanalysis, huge sample-to-sample variability was observed from single cells amplified by MALBAC, suggesting that further independent evaluation of the method may be required to assess its performance.

Detection of Somatic Aneuploidy and CNVs from Normal Tissues

Somatic aneuploidy and CNVs are most widely studied in the setting of cancers, via either traditional SNP-array from bulk samples or sequencing-based single-cell analysis (Beroukhim et al. 2010; Navin et al. 2011; Jacobs et al. 2012; Laurie et al. 2012). Bulk analysis allows for the characterization of the prevalence of different types of copy number alternations, such as segmental CNVs and chromosomal arm-level copy number alternations; previous bulk analyses have shown that segmental CNVs are often biased towards smaller sizes with a median size of 1.8Mb, whereas chromosomal arm-level CNVs are 30-fold more frequent than segmental CNVs at the matched size (Beroukhim et al. 2010). On the other hand, single-cell copy number analysis of tumor samples with high grade CNVs revealed multiple clonal populations with distinct copy number profiles, allowing for the tracking of tumor cell lineages (Navin et al. 2011). In contrast to the high prevalence of somatic CNVs affecting approximately ~30% of the genome in cancerous cells (Beroukhim et al. 2010), large CNVs are must less well-tolerated by normal cells. A recent study of tissue-specific somatic CNVs from normal individuals has revealed that the majority of events are below 50kb in size and show an average frequency between 2-6 events per tissue, accounting for only a negligible fraction of the genome (O'Huallachain et al. 2012; O'Huallachain et al. 2013). In addition, somatic aneuploidy at the chromosomal arm-level was not seen. Larger somatic CNVs (50kb to whole chromosome arm) can be detected from normal individuals at low mosaic states no greater than 1% on average; and the frequency increases with

age from 0.23% under 50 years to 1.91% between 75 and 79 years (Jacobs et al. 2012; Laurie et al. 2012). Compared to the low prevalence of somatic CNVs in terminally differentiated tissues, chromosomal abnormality seems to be much more common during early embryonic development. Large unbalanced chromosomal rearrangements were detected at frequency ranging from 30-80% of embryos at the preimplantation blastomere stage by a number of single-cell copy number studies (Vanneste et al. 2009; van Echten-Arends et al. 2011; Yin et al. 2013). Such a high frequency suggests that early embryonic cell divisions are particularly error-prone with a high tendency towards chromosome missegregation and rearrangement; however, most of the abnormal cells seems to be under negative selection during further embryonic development, resulting in a much lower frequency of chromosomal abnormalities in adults.

Somatic aneuploidy has been proposed as a mechanism in generating genetic diversity among post-mitotic neurons (See details in Chapter 1 *Somatic variants in normal brains*). However, based on our knowledge from studies on other tissues, we expect a rather low rate of aneuploid cells developing into neurons due to their expected proliferative disadvantages, unless there is an acquired mechanism to generate aneuploidies and/or unbalanced chromosomal rearrangments only at the terminal differentiation of neurons. Current evidence about somatic neuronal aneuploidies is solely based on interphase FISH probing a single or small number of chromosomes. Estimates of aneuploidy frequencies from different studies are also highly variable, ranging from 1.3%-40%, and rates seem to increase with age (Faggioli et al. 2011). Interphase FISH suffers from high false positive rates due to non-specific priming of the probes and offers an incomplete picture of the copy number state of the entire genome; therefore, more rigorious single-cell genome-wide copy number profiling would be necessary to assess previous observations and to potentially explaining experimental discrepencies.

Results

Amplification linearity of single cell genomes

Although we have shown in Chapter 3 that MDA reliably amplifies single cell genome to a coverage ~90%, and is able to detect aneuploid single neurons with trisomy 18 (**Figure 3-7**), the stochastic amplification noise spanning $\pm$ 1 copy number significantly increases the false positive rate and prevents us from studying sub-chromosomal segmental CNVs with this method (**Figure 3-8A, C**). We therefore explored an alternative single-cell whole-genome amplification method, GenomePlex WGA4, which was shown to amplify the single-cell genome more linearly albeit at significantly lower coverage (<10%) (Navin et al. 2011). To be able to compare the two methods directly, we developed a QC metric adapted from Affymetrix MAPD (Median Absolute Pairwise Difference) algorithm to measure the stochastic amplification noise of the two methods. MAPD measures the absolute differences between the copy number ratios of each adjacent bins, and then takes the median. This algorithm has been widely used as a quality assessment of SNP Array 6.0 copy number data, and a MAPD score >0.4 typically associates with poor quality copy number calls and high false positive rates (Affymetrix, Inc. 2008). We first measured the MAPD scores of 4 single-cell MDA samples, compared with 1 100-cell MDA sample together with two bulk unamplified samples from the same individual at different read depths to determine the effect of sequencing read depth on copy number data quality (**Figure 4-1A**). We found that MDA amplified samples give significantly higher MAPD scores compared to unamplified bulk samples, reflecting the stochastic amplification noise associated with MDA. A MDA amplified 100-cell sample give a lower MAPD score than single cell samples, suggesting that the amplification of the single genome introduces another layer of amplification noise on top of MDA. Notably, the sequencing read depth, measured by number of reads per genomic bin, has

little effect on MAPD scores of MDA samples. In contrast, increased read depth helped to reduce the MAPD scores of bulk samples (**Figure 4-1A**). These data suggest that the copy number noise in MDA amplified samples is due to the genome amplification, instead of insufficient read depth for copy number profiling. The most cost effective read depth (~750 reads/bin, 4.5 millions/sample, 0.07X genome coverage) was then chosen for following experiments. At this read depth, we are able to safely multiplex 32 single cell samples into 1 Illumina HiSeq lane for copy number analysis at ~500kb resolution.

We then compared the amplification noise of GenomePlex WGA4 amplified single cells to MDA amplified single cells. Since we suspect that the tissue type and quality of the original sample may affect genome amplification quality, the comparison was limited to single cells derived from the same tissue source (UMB4643-cortex). Indeed, the MAPD scores of WGA4 amplified samples were consistently lower than MDA amplified samples (**Figure 4-1B**). This result is consistent with the observed tighter distribution of copy number ratios by WGA4 compared to MDA (**Figure 4-1C**). Therefore, we concluded that MAPD score is a reliable measurement of data quality of single-cell genome amplified samples, and that WGA4 is a better method for single-cell copy number profiling. We further compared the data quality of the two methods at different genomic bin sizes, including ~500kb, ~150kb and ~60kb, with normalized read counts for each bin. Since we didn't sequence our WGA4 samples to high-enough read depth for this analysis, we used data from Navin et al. (2011) on 4 wildtype control single cells amplified by WGA4 to compare with our 4 wildtype cortical neurons amplified by MDA. We showed that with decreased bin sizes, MAPD scores increased for both amplification methods, suggesting that both methods introduce greater amplification noise at smaller local regions (**Figure 4-2A**). This is expected because larger bins contain many independently amplified

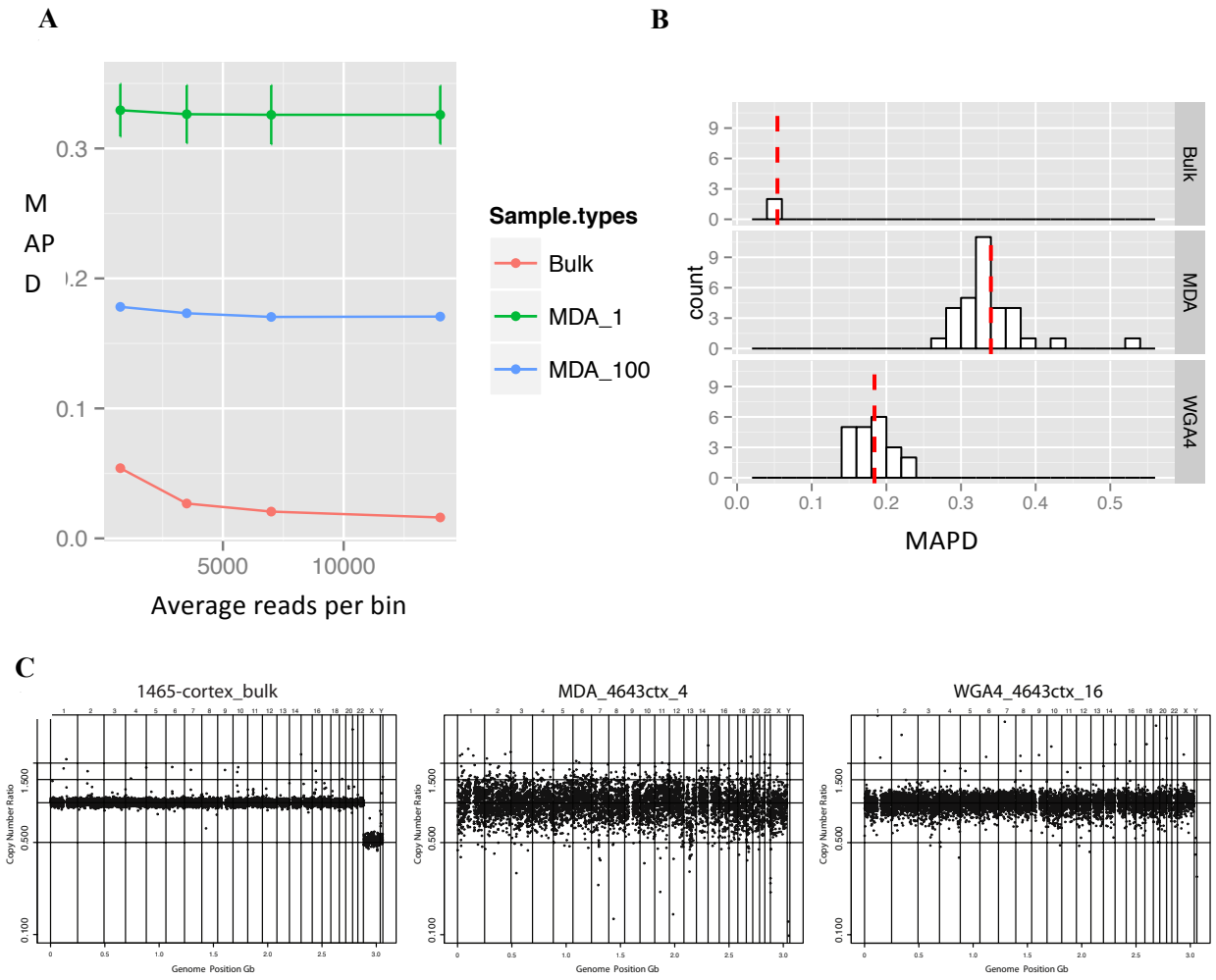


Figure 4-1. Comparison of amplification linearity between MDA and WGA4 by MAPD at ~500kb bin size.

(A) MAPD score of 1-cell and 100-cell MDA amplified samples are various read depth compared to bulk samples. Increase of read depth (total number of reads per bin) does not bring down the MAPD score of MDA samples (N=4 for MDA single cells "MDA_1"; N=1 for MDA 100-cell "MDA_100"; N=2 for bulk; error bar = ± SD). X-axis represents the number of reads per bin and the Y-axis represents the MAPD score.

(B) Histogram of MAPD distribution of single neurons from 4643 cortex, amplified by either MDA or WGA4. Bulk DNA from 1465 cortex is used to define the baseline of MAPD (N=21 for WGA4; N=32 for MDA; N=2 for bulk). WGA4 amplified samples consistently have smaller MAPD, with average at 0.184±0.026, compared to MDA amplified samples, with average at 0.340±0.047. There are two outlier cells from MDA amplified samples with high MAPD score. Average reads per bin is 734±14 for bulk, 735±118 for MDA and 313±40 for WGA4 samples (Error = ± SD). The red dash lines denote the mean MAPD scores of each method.

(C) Copy number profiling at 6,000 bins (~500kb bin size) of representative samples from Figure 4-1B. Bulk, MDA amplified single neuron and WGA4 amplified single neuron are shown respectively from left to right. MDA sample has the highest noise spanning +/- 1 copy number; copy number distribution of WGA4 samples is tighter compared to MDA, spanning +/- ~0.5 copy. Y-axis denotes the copy number ratio (CNR) respect to euploid genome. Therefore copy number ratio 1 stands for copy number 2, 0.5 stands for copy number 1 and 1.5 stands for copy number 3.

fragments; therefore, over or under amplification at a small region within a large genomic bin is unlikely to change its overall copy number; whereas these stochastic nonlinearly amplified fragments are more likely to influence the copy number of smaller genomic bins which includes fewer independent fragments. Notably, the MAPD score of bulk samples do not change with bin sizes, confirming that the increase of MAPD in amplified samples is not due to binning artifacts. Furthermore, although MDA samples always give higher MAPD score and higher amplification noise, this defect can be corrected by increasing genomic bin size at the cost of copy number resolution. The MAPD score of MDA samples at ~500kb bin size is similar to WGA4 samples at ~60kb bin size (0.33$\pm$0.02 and 0.33$\pm$0.01, respectively; error = $\pm$SD); therefore, the choice of ~500kb bin size for copy number profiling on MDA amplified samples is adequate. With this approach, we should reliably detect copy number changes at megabase scale from MDA amplified single cells, evident by the successful detection of trisomy 18 from single neurons in our earlier experiments (**Figure 3-7B**). Since we used WGA4 data from others (Navin et al., 2011) for the comparisons in **Figure 4-2A**, we also compared the data quality of our WGA4 samples from cortical neurons and cultured lymphocytes with cancer and normal samples from Navin et al. (2011) to show that the MAPD scores from all 4 tissue types are comparable, with slightly higher average and wider distribution in our samples (**Figure 4-2B**). These data also demonstrate that MAPD score only measures stochastic data noise, but is not altered by biological CNV states since the cancer samples with high copy number changes have the same average MAPD score with its normal control (compare "cancer" and "control"). Therefore, MAPD is a superior quantitative measurement of the overall linearity of genome amplification compared to the measurement of "uniformity" on the amplified genomes used in another study (Zong et al. 2012). We also noticed that three outlier cells from the lymphocytes samples,

presumably due to poor amplification, can be readily differentiated by their increased MAPD; therefore, in addition to measuring data quality across different methods, MAPD can also used as QC metrics to exclude poor quality samples from copy number analysis (discussed in the following section).

The other challenge in copy number profiling on single-cell amplified genomes is locus dropouts, in which regions fail to amplify or amplify poorly in a stochastic manner. These regions, if large enough, will be falsely recognized as genomic deletions of either one or both copies. We therefore quantified the fraction of the genome that "dropped out" by 1 and 2 copies at ~500kb bin size. For MDA-amplified single neurons from 4643 cortex, a median of 1.23% of the genome is dropped out by 1 copy and 0.23% of the genome is dropped out by 2 copies (**Figure 4-3**). For WGA4-amplified single neurons, a median of 0.32% of the genome is dropped out by 1 copy and 0% of the genome is dropped out by 2 copies (**Figure 4-3**). In contrast, zero bins were dropped out as either 1 or 2 copies from bulk samples. The median is used to quantify dropout rates because there are outlier cells in both sample sets that skew the mean. Since both sample sets were derived from the same tissue with the same germline genotype, we can conclude that the increased dropout rate seen in MDA samples compared to WGA4 is due to MDA amplification. However, the 0.32% of the genome dropped out by 1 copy in WGA4 samples consists of a mixture of true somatic heterozygous deletions and artifactual locus dropout due to WGA4 amplification. On the other hand, the 0% dropout by 2 copies from WGA4 samples holds great promise on detecting somatic homozygous deletions at megabase scale. In addition, we found outlier cells with large number of dropouts from both MDA and WGA4 amplified samples (**Figure 4-3**); the biological significance of these cells will be further discussed in the following sections.

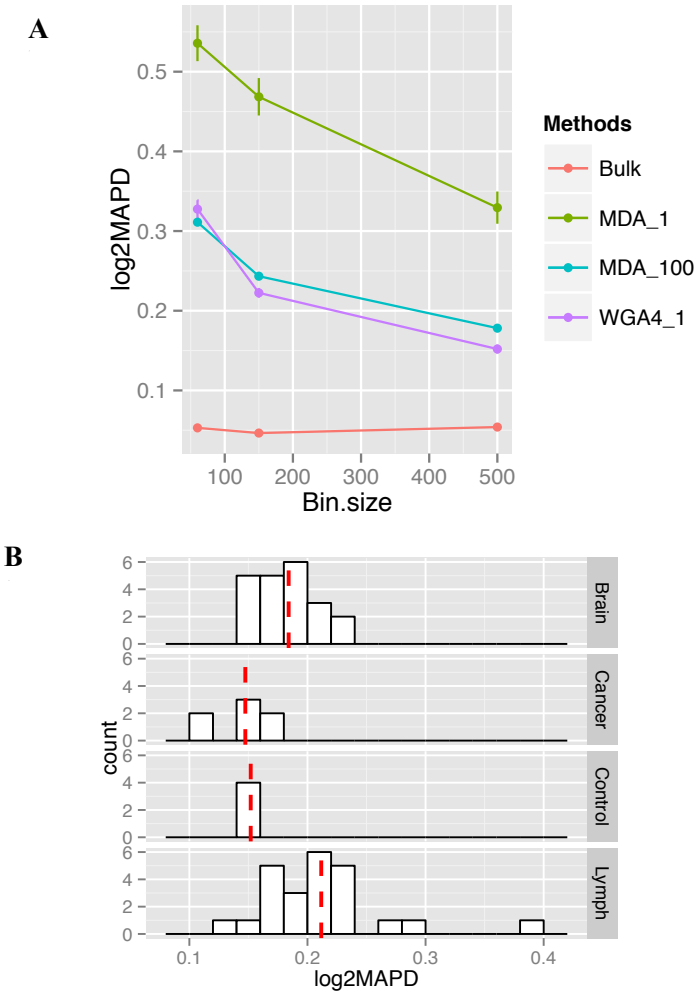


Figure 4-2. Effects of bin sizes on MAPD scores of both methods at normalized read depth.
(A) Average MAPD score of bulk, MDA amplified and WGA4 amplified single cell samples are plotted against various bin sizes, including 500kb (6,000 bins in total), 150kb (20,000 bins in total) and 60kb (50,000 bins in total). MAPD scores of both MDA and WGA4 amplified samples increase as bin size get smaller, whereas the MAPD score of bulk sample remains unchanged with changing bin sizes, suggesting that both amplifications introduce more prominent noise at smaller local regions (N=2 for bulk; N=4 for MDA single cells "MDA_1"; N=1 for MDA 100-cell "MDA_100"; N=4 for WGA4 single cells "WGA4_1"; error bar = ± SD).

(B) Histogram of MAPD distribution of single cells from different tissue types amplified by WGA4. 7 breast cancer single cells ("cancer") and 4 normal control cells for the cancer study ("control") were from Navin. et al. (2011). 24 single neurons from 4643 cortex ("Brain") and 24 cultured wildype lymphocytes ("Lymph") are from our study. MAPD scores are generally comparable among all 4 tissue types, with slightly wider distribution of lymphocyte samples. Average MAPD score of the 4 tissue types are 0.184±0.026, 0.152±0.006, 0.147±0.026 and 0.212±0.053 respectively from top to bottom (error = ±SD). The average read depth per bin of the 4 tissue types are 313±40, 275±34, 451±176 and 258±30 respectively from top to bottom (error = ±SD).

In conclusion, we developed a QC metrics to assess the single-cell genome amplification linearity of two independent methods and concluded that both methods are suitable for copy number analysis of single cells at the chromosomal level whereas only WGA4 should be used for segmental CNVs analysis. The QC metric based on MAPD score reliably measures the stochastic noise of copy number profiles and is not altered by biological CNVs states of the sample. Therefore, it can be further applied to all single cell samples as a quality control assessment to exclude poor quality cells from further analysis to minimize the effect of amplification bias on the estimation of somatic CNV prevalence.

Chromosomal copy number analysis of single neurons from normal human brains

Following out analysis of CNV quality, both MDA and WGA4 amplified samples were used to study the prevalence of somatic aneuploidy in wildtype cortical neurons. A total of 139 single cells were analyzed, including 97 single cortical neurons from 3 normal adults (UMB1465, UMB4638 and UMB 4643), 18 single neurons from a trisomy 18 fetus (UMB866) and 24 cultured single lymphocytes derived from a normal adult GM21781 (**Table 4-1**). With a QC threshold of MAPD <= 0.45, 82 single neurons from normal individuals, 9 single neurons from the trisomy 18 individual and 24 single, wildtype lymphocytes were included in the chromosomal copy number analysis. All WGA4 amplified samples (26 single neurons and 24 single lymphocytes) passed the QC threshold; however, the sample qualities of MDA amplified samples are highly variable, depending on the individual (**Figure 4-4A**). We suspect that the quality of the tissue source could be the cause of the variation; this highlights the importance of applying a QC metric for all amplified single cell samples prior to any copy number analysis. In addition, 4 samples that failed the initial multiplex PCR screen were sequenced and included in

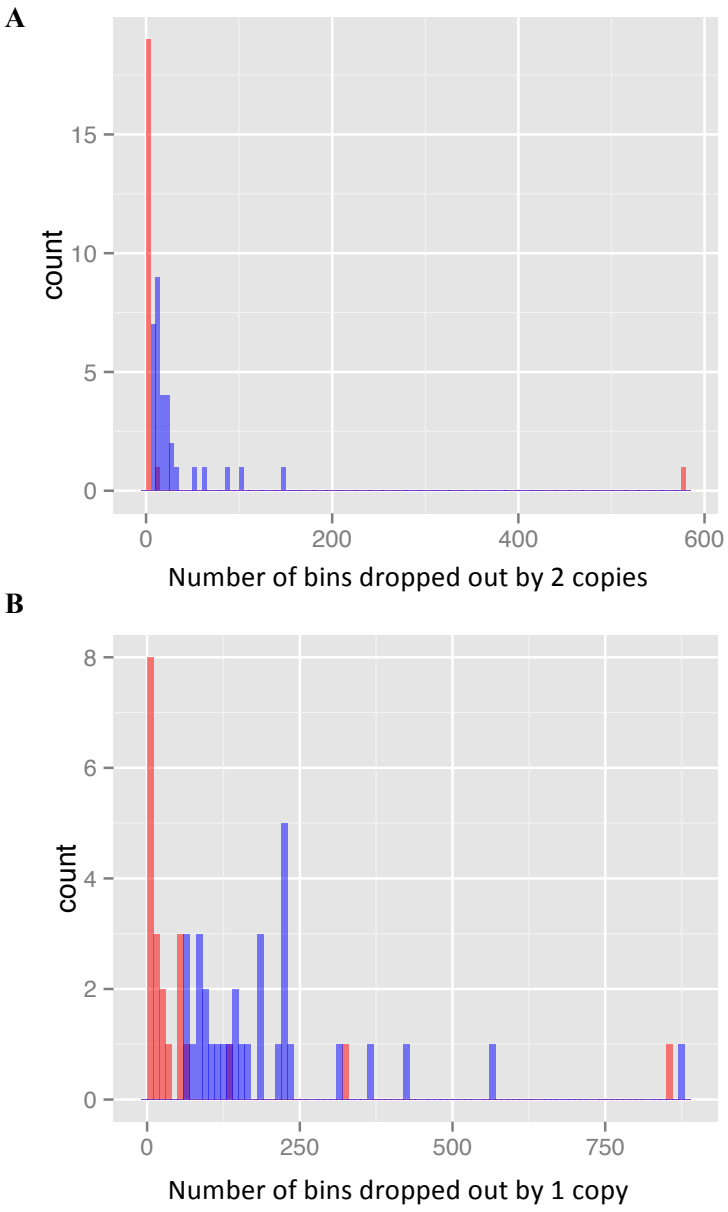


Figure 4-3. Locus dropout of MDA and WGA4 amplified single cells at ~500kb bin size.
(A) Histogram of number of bins with Log2CNR < -2 of single neurons from 4643 cortex, amplified by either MDA (blue) or WGA4 (red) (N=21 for WGA4; N=32 for MDA). The median 2 copy dropout of WGA4 samples are 0 out of 6000 bins, with 1 outlier sample of >500 bins dropped out. The median 2 copy dropout of MDA samples are 13.5 out of 6000 bins, contributing to ~0.23% of the genome.

(B) Histogram of number of bins with Log2CNR < -1 of single neurons from 4643 cortex, amplified by either MDA (blue) or WGA4 (red) (N=21 for WGA4; N=32 for MDA). The median 1 copy dropout of WGA4 samples are 19 out of 6000 bins, contributing to 0.32% of the genome, with 1 outlier sample of >800 bins dropped out. The median 1 copy dropout of MDA samples are 73.5 out of 6000 bins, contributing to ~1.23% of the genome.

Table 4-1. Summary of all single cells analyzed for chromosomal copy numbers.

Individual	Karyotype	Tissue type	Cell type	Amplification method	# of cells analyzed	# of cells passed QC (MAPD <= 0.45)	# of euploid cells	# of aneuploid cells
UMB1465	46XY	Cortex	Neuron	MDA	7*	6	6	0
UMB4638	46XX	Cortex	Neuron	MDA	32	20	19	1
UMB4643	46XX	Cortex	Neuron	MDA	32	30	30	0
UMB4643	46XX	Cortex	Neuron	WGA4	26	26	24	2
UMB866	47XY, 18	Cortex	Neuron	MDA	18	9	0	9**
GM21781	46XY	Cultured lymphocytes	Lymphocyte	WGA4	24	24	24	0
				TOTAL	139	115	103	12

* 4 of these single cells were sequenced at >30X coverage

** trisomy 18

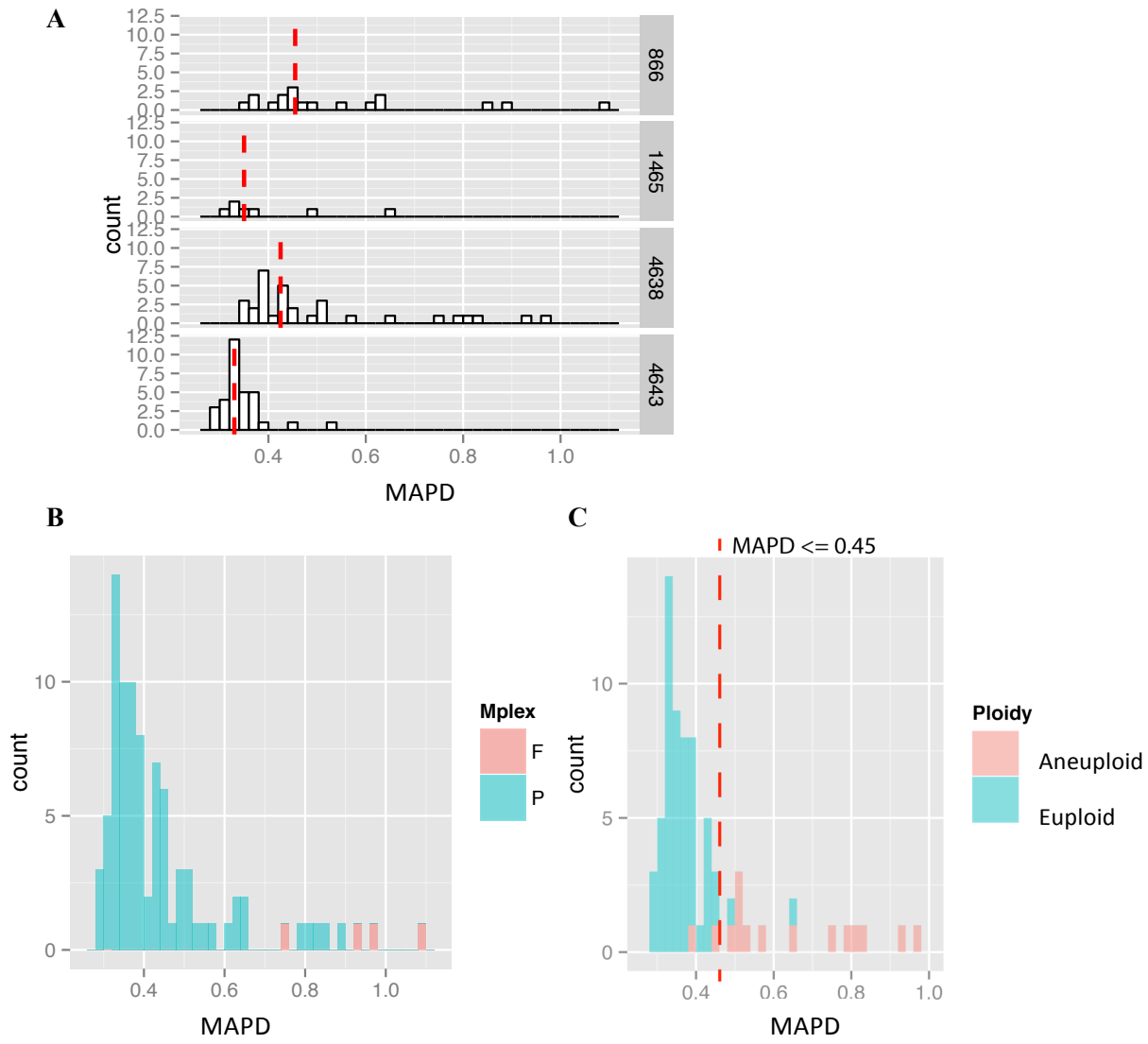


Figure 4-4. MAPD QC metric for MDA single-cell chromosomal copy number analysis.
(A) Histogram of the MAPD scores of MDA-amplified single cells from 4 different tissue sources. N-18 for UMB866 cortex; N=7 for UMB 1465 cortex; N=32 for UMB 4638 cortex and N=32 for UMB4343 cortex. The red dash lines denote the mean MAPD score of each tissue sources.

(B) Histogram of the MAPD scores of MDA-amplified single cells from all 4 individuals either passed ("P", green) or failed ("F", red) the initial multiplex PCR screen. N=86 for passed and N=3 for failed multiplex PCRs. All samples that failed the multiplex PCR appeared as poor quality outliers based on their MAPD scores.

(C) Histogram of the MAPD scores of MDA-amplified single cells from all 3 normal individuals either called as "euploid" (green) or aneuploid (red) prior to QC filtering. N=71 in total; N=56 for "euploid" and N=15 for "aneuploid". With a QC threshold of MAPD <= 0.45, 56 samples passed the threshold and 15 samples failed. Among the samples passed the threshold, 54/56 are "euploid" and 2/56 are "aneuploid". Among the samples failed the threshold, 2/15 are "euploid" and 13/15 are "aneuploid." The red dash line denotes the QC threshold at MAPD <= 0.45.

the analysis to confirm whether the multiplex PCR helps to distinguish poorly amplified samples. Indeed, all samples that failed the multiplex PCR appear as outliers with high MAPD scores (**Figure 4-4B**); however, multiplex PCR is not sufficient to filter out all poor-quality samples since a few samples that passed the multiplex PCR screen also appeared as outliers with high MAPD scores, further suggesting that additional QC is necessary for copy number analysis. We relaxed the QC threshold of MAPD from 0.4 to 0.45 to include more cells for the analysis with minimal loss of specificity.

Among the 71 MDA amplified single neuron samples from normal individuals, 56 of them passed the QC threshold, and 2 out of the 56 neurons have aberrant chromosomal copy numbers (**Figure 4-4C**). On the other hand, 13/15 samples that failed the QC appeared to have noisy copy number profile with multiple chromosome gains and losses, confirming that the MAPD is a powerful tool on differentiating false positive copy number aberration caused by poor amplification. 9 out 18 trisomy 18 cells passed the QC, with an average copy number of 3.01 ± 0.17 at chromosome 18, demonstrating the high sensitivity of method on detecting copy number changes at the chromosomal level from single cells amplified by MDA (**Table 4-1** and **Figure 4-5B**). The two cells with copy number aberrations are MDA_4638ctx_24 and MDA_4643ctx_14 (**Figure 4-5C, D**, respectively). MDA_4638ctx_24 has a grossly aberrant genome with copy numbers alternating between 0, 1 and 2 (**Figure 4-5E**). Multiple chromosomes (Chr2, Chr6, Chr11 and ChrX) or chromosome arms (Chr10q, Chr12q, Chr21p) have a discrete copy number 1, and additional chromosomes have copy numbers alternating between 0 and 1 (Chr3, Chr4p, Chr5, Chr17) (**Figure 4-6**). The other aberrant neuron, MDA_4643ctx_14 has a noisy copy number profile overall with a borderline MAPD score at 0.44 and an intermediate copy number of chromosome X at 1.4 (**Figure 4-5D**). The copy

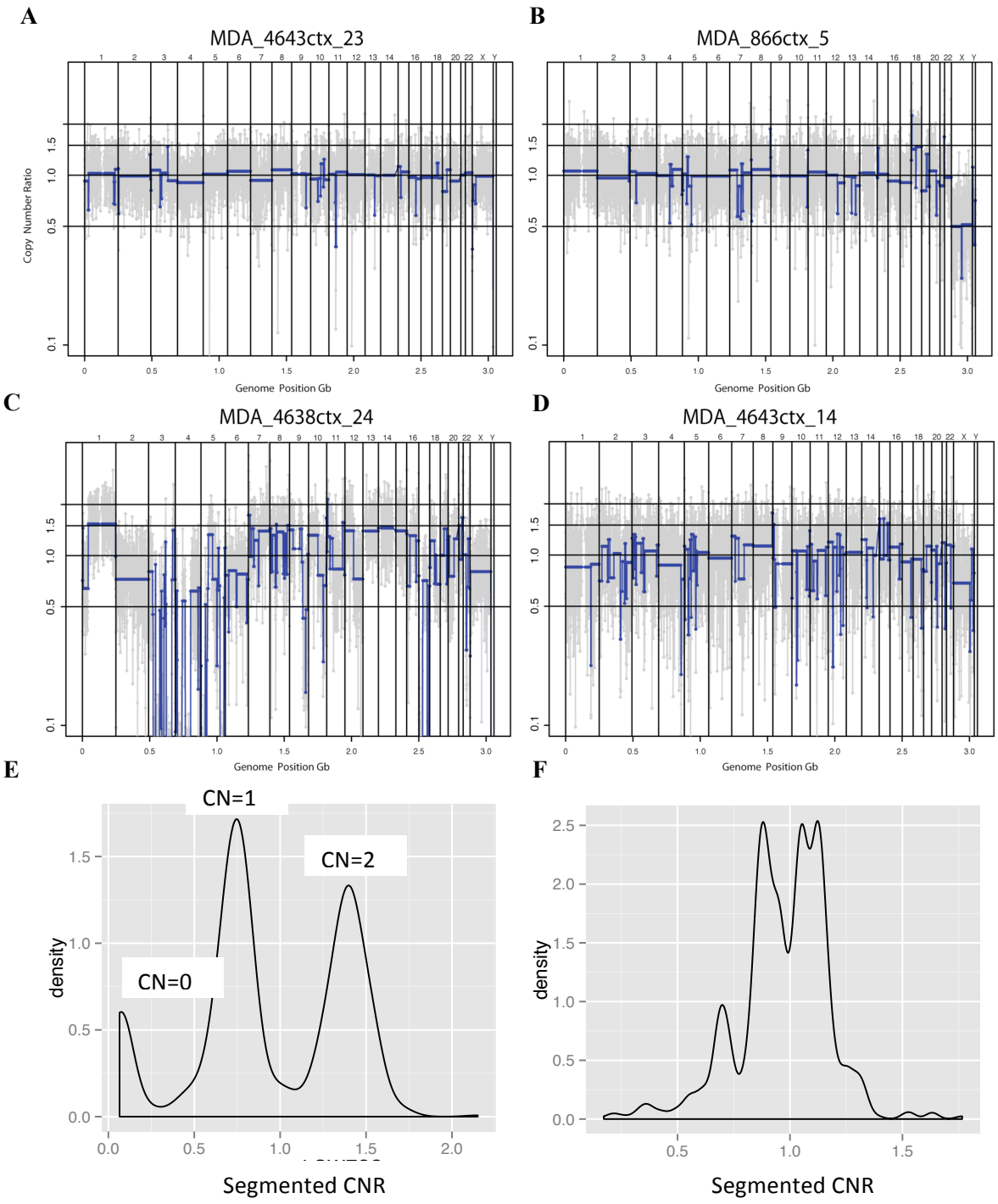


Figure 4-5. Copy number profiles of MDA amplified single neurons at ~500kb bin size.
(A-D) Genome-wide copy number profiles with segmentation of 4 representative single neurons. **(A)** represents an euploid neuron; **(B)** represents a positive control neuron with copy number gain at Chr18; **(C)** represents a neuron with grossly aberrant genome and **(D)** represents the other neuron called as "aneuploid" with a intermediate copy number at ChrX (CNR=0.7, CN=1.4). X-axis represents all 6,000 genomic bins arranger by their genomic positions by chromosomes. Log2 transformed Y-axis denotes the copy number ratio (CNR) respect to euploid genome.

(E-F) Copy number density plot of the two "aneuploid" single neurons in **Figure 4-5C-D**, respectively, Neuron MDA_4638ctx_24 showed clear discrete copy numbers at CN=0, 1, 2; whereas neuron MDA_4643ctx_14 showed a major peak around CNR=1 and an small peak representing ChrX between CNR=0.5 and 1.0. X-axis represents the segmented mean copy number ratio (CNR).

number ratio density plot showed a major peak around ratio 1.0 and a small peak at 0.7, consistent with the predicted copy number of chromosome X at 1.4 (**Figure 4-5F**). Based on the trisomy 18 samples, the standard deviation of altered copy number is 0.17; therefore, copy number 1.4 at chromosome X of a female cannot be considered as a full copy number loss, and so may be a technical artifact.

24 single lymphocytes and 26 single neurons amplified by WGA4 were also analyzed for chromosomal copy number. All 50 samples passed the QC threshold of 0.45 although the distribution of MAPD score of the WGA4 samples showed 3 outlier samples from GM21781 lymphocytes (**Figure 4-2B**). However, despite their significantly nosier copy number profiles with increased number of copy number changes at a sub-chromosomal scale (Compare **Figure 4-7B** to **Figure 4-7A**), all 3 outlier samples are still called as euploid in the copy number analysis, further confirming that the MAPD score faithfully reflects the data quality and the QC threshold used in the chromosomal copy number analysis is adequate. On the other hand, it also suggests that for segmental CNV analysis, a more stringent QC threshold would be necessary. 2 out of 26 neurons from 4643-cortex (WGA4_4643ctx_30 and WGA4_4643ctx_22) exhibit alternated copy numbers at multiple chromosomes (**Figure 4-7C, D**, respectively). WGA4_4643ctx_30 exhibits grossly imbalanced chromosomal number, with copy numbers alternating between 0, 1 and 2, similar to what has been observed for MDA_4638ctx_24 (**Figure 4-7C, Figure 4-8C** and **Figure 4-5C, E**). Multiple chromosomes exhibited copy numbers ranging from 0 to 1 (Chr14, 15, 16 and ChrX) and additional chromosomes exhibited alternating copy numbers from 1 to 2 (Chr1, Chr7, Chr9, Chr11, Chr17, Chr19) (**Figure 4-7C** and **Figure 4-8C**). WGA4_4643ctx_22 exhibited 4 copy number states, with multiple chromosomes exhibiting copy number gains (Chr4, Chr7, Chr16, Chr7 and ChrX) and multiple chromosomes with alternating

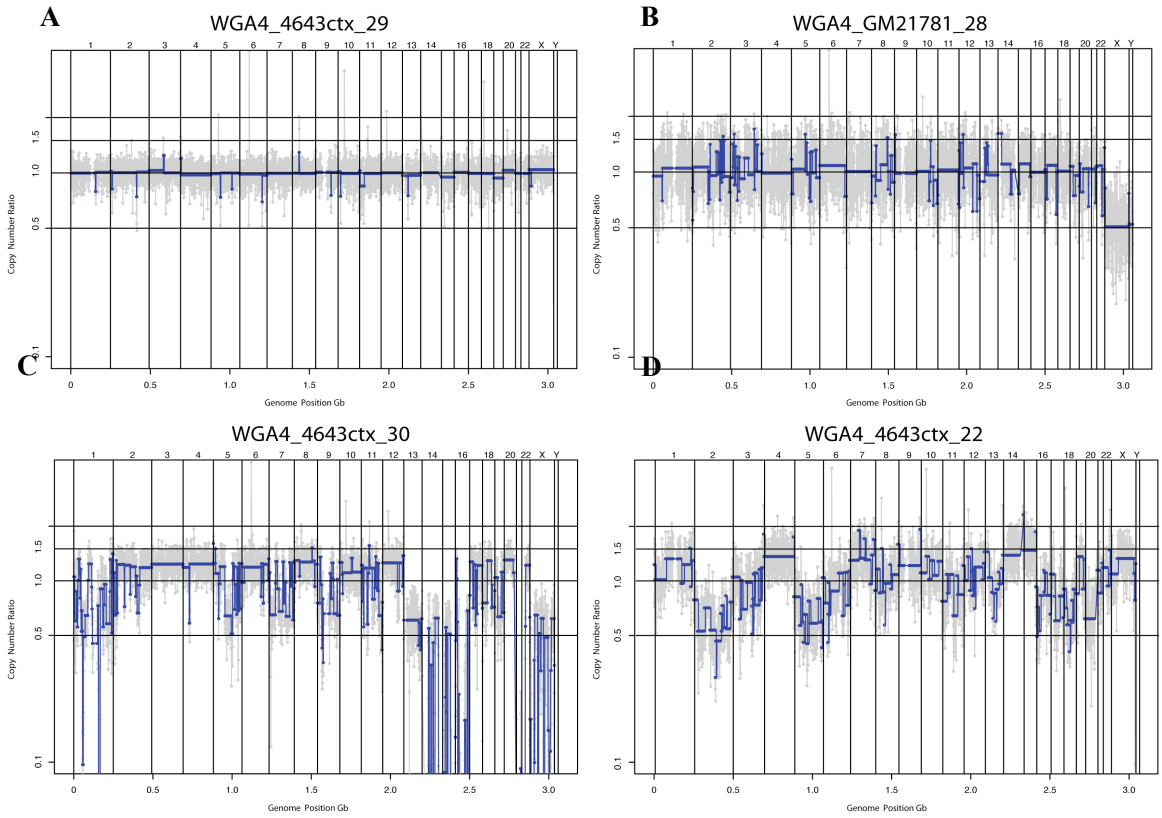


Figure 4-7. Copy number profiles of WGA4 amplified single cells at ~500kb bin size.
Genome-wide copy number profiles with segmentation of 4 representative single cells.

(A) represents an euploid neuron with low MAPD score (MAPD=0.15);

(B) represents a euploid single lymphocytes with a MAPD score appeared as an outlier among WGA4 samples (MAPD=0.27);

(C) represents a neuron with grossly aberrant genome and large homozygous deletions; and

(D) represents the other neuron called with grossly aberrant genome and no homozygous deletions, but copy number gains at chromosomal level.

X-axis represents all 6,000 genomic bins arranger by their genomic positions by chromosomes. Log2 transformed Y-axis denotes the copy number ratio (CNR) respect to euploid genome.

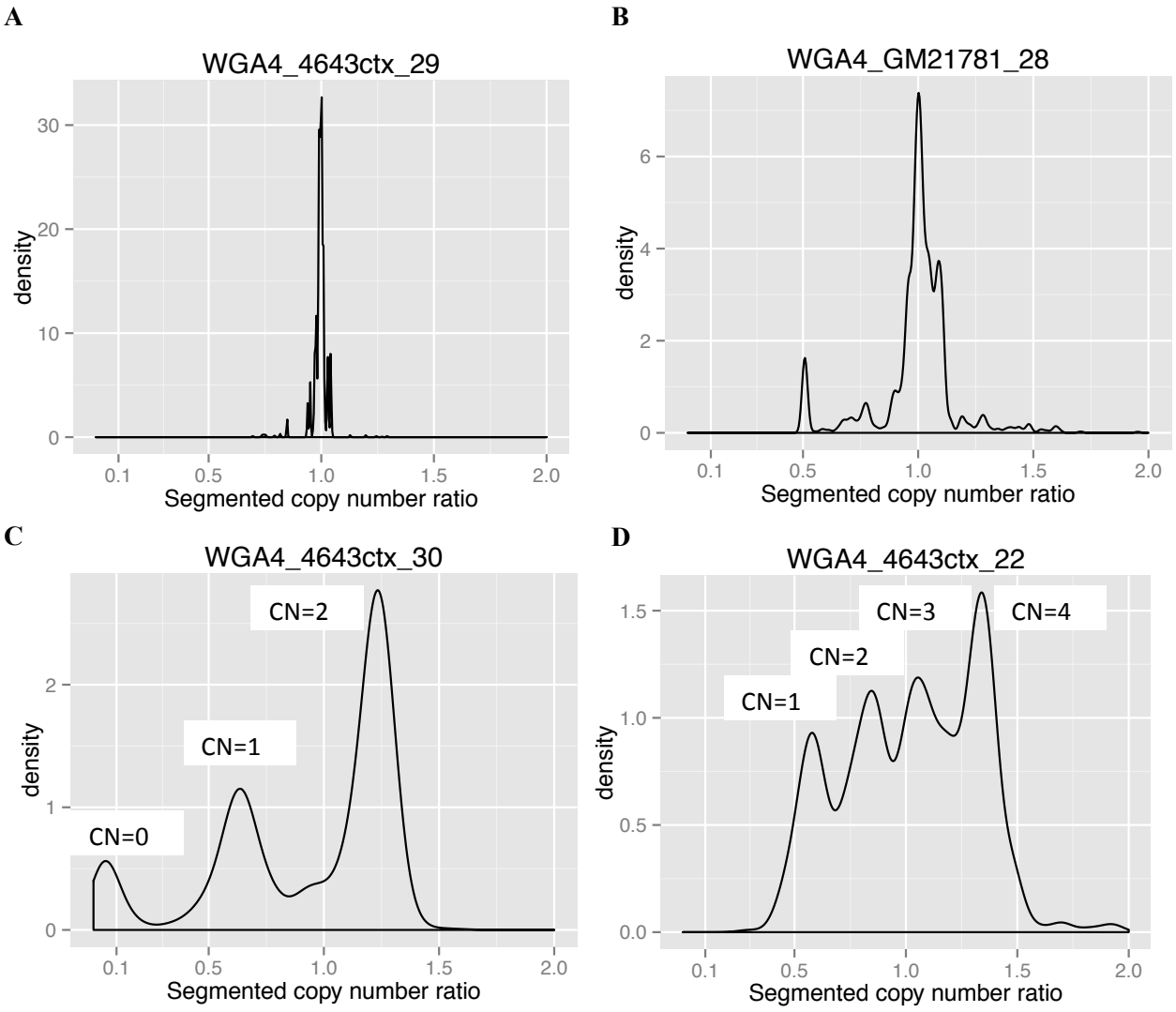


Figure 4-8. Copy number ratio density plot of WGA4 amplified single cells at ~500kb bin size. Copy number ratio density plot of the four single cells presented in **Figure 4-7**, respectively.

(A-B) Euploid single cells showed single peak at CNR=1 and additional peak at CNR=0.5 for male sample (GM21781).

(C-D) Neuron WGA4_4643ctx_30 showed clear discrete copy numbers at CN=0, 1, 2; and neuron WGA4_4643ctx_22 showed 4 overlapping peaks presumably representing copy numbers at CN=1, 2, 3, 4.

X-axis represents the segmented mean copy number ratio (CNR).

copy number between 1 and 2 (**Figure 4-7D** and **Figure 4-8D**). However, WGA4_4643ctx_22 did not show loss of both copies of any chromosomes as observed in the other two aberrant cells.

In conclusion, we analyzed 82 single cortical neurons from 3 normal individuals with a quality control threshold at MAPD <= 0.45. 79 out 82 neurons are euploid, accounting for 96.3% of the population (95% confidence interval: 89.8%-98.8%). We did not identify any aneuploid neuron with discrete copy number change of a single chromosome. This result puts an upper bound at 4.5% for the prevalence of aneuploid neurons with single chromosome gain or loss in normal individuals. Instead of detecting canonical aneuploid neurons as expected, we identified 3 aberrant neurons with genome-wide copy number imbalances. This observation is unlikely to be a technical artifact specific to a certain genome amplification method since we identified similar aberrant cells from both MDA and WGA4 amplified samples. This observation is also not limited to a specific individual as we identified them from both individuals with more than 10 single neurons analyzed. However, we cannot rule out other technical reasons that may lead to the observation of these cells. Further discussion on the possible causes of these aberrant cells is carried out in the Discussion section.

Segmental CNVs of single neurons from normal human brains

Segmental CNVs are defined as copy number gains and losses at sub-chromosomal scale. Although they can also be large in size, their molecular mechanism is distinct from aneuploidy. Large CNVs, up to a few megabases, are also one of the most common genetic causes of neurodevelopmental disorders (Sanders et al. 2011; F. Zhang et al. 2009). We performed genome-wide copy number profiling on the 24 euploid single neurons amplified by WGA4, to identify somatic CNV candidates down to 2 megabase in size. The four different CNV states

(CNV0, CNV1, CNV3 and CNV4) were analyzed separately across all the samples since the specificity and sensitivity of each CNV state is different.

As discussed in the previous section, additional quality control assessments are applied to the WGA4 amplified samples for segmental CNV analysis. In addition to the amplification noise measured by MAPD, locus dropout introduced by nonlinear amplification can also lead to false positive CNV calls. Since all these single neurons are derived from the same tissue, we expect the CNV states of them to be similar, potentially with a low level of somatic variations. We measured the number of bins at ~500kb size that were dropped out by copy number 1 (log2GCNR < -1) across all samples and found the median to be 26.5; however, a few samples exhibit as outliers with > 100 bins dropped out by copy number 1 (**Figure 4-9A**). We therefore conclude that such increase in dropout rate is unlikely to be physiological but more likely to be a technical artifact caused by the whole genome amplification. Samples with >100 apparent CNV 1 were then excluded from further segmental CNV analysis. In fact, the 6 samples excluded have the 6 highest MAPD scores among all 24 samples, supporting the interpretation that the excluded samples have relatively poor quality compared to the rest, and again demonstrating the direct correlation between MAPD score and sample quality in multiple aspects (**Figure 4-9B**).

18 single cortical neurons that passed the dropout quality control were then included in the segmental CNV analysis, which identifies genomic segments with at least 5 consecutive genomic bins and a size larger than 2Mb. For copy number gains, 0 candidate were identified at CNV4, and 5 candidates were identified at CNV3 (**Table 4-2**, **Appendix Table 4-1**). Since mapping and binning artifacts can lead to false positive copy number gains, the candidates were then BLATed using UCSC genome browser. 3 out 5 candidates are BLATed to centromeres,

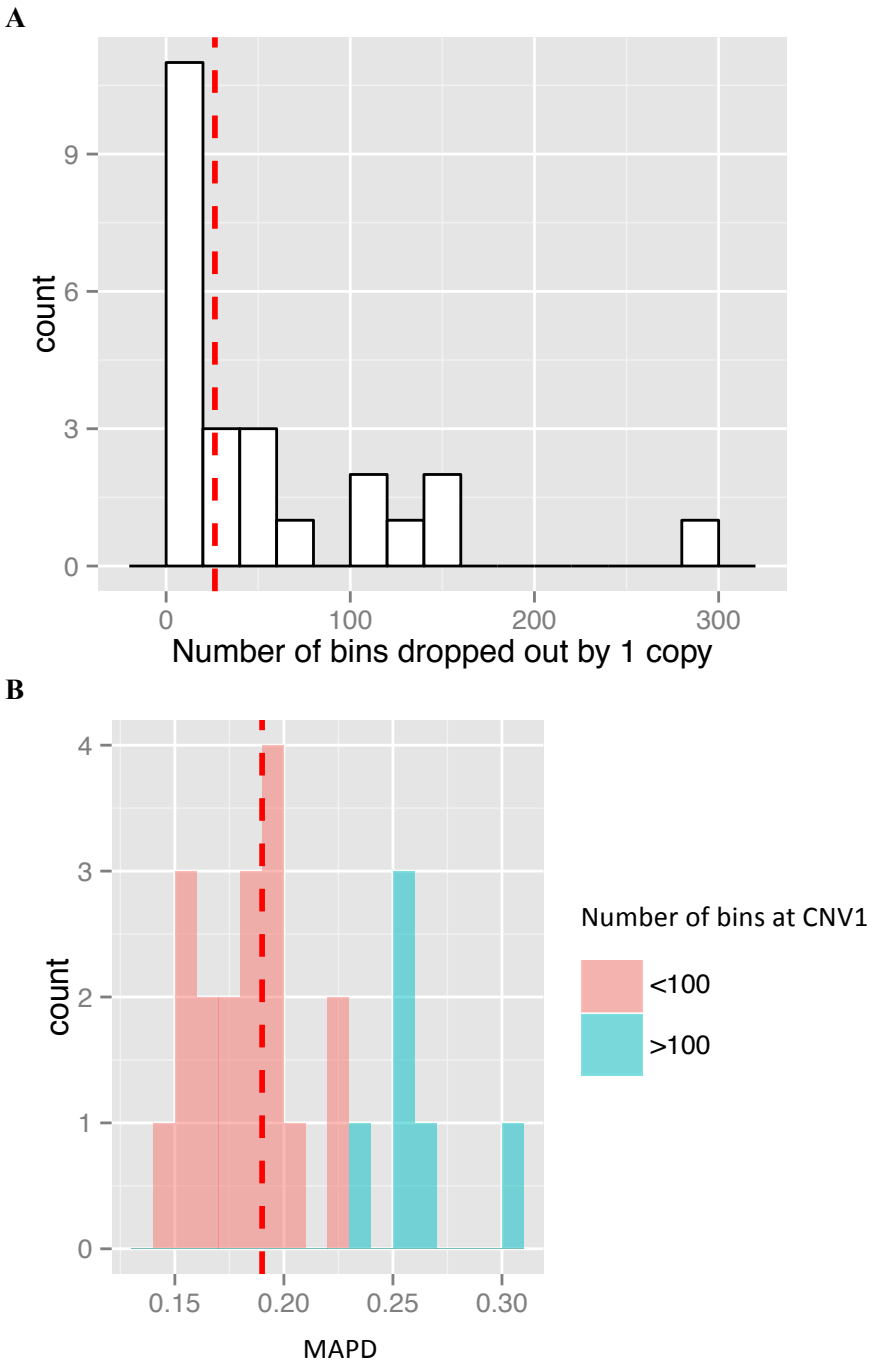


Figure 4-9. Locus dropout by copy number 1 of WGA4 amplified single neurons at ~500kb bin size.

(A) Histogram of number of bins with Log2CNR < -1 of euploid single neurons from 4643 cortex, amplified WGA4 (N=26). The median 1 copy dropout of WGA4 samples are 26.5 out of 6000 bins, with several outlier samples of >100 bins dropped out. Red dash line denotes the median MAPD score of all 24 samples.

(B) Histogram of the MAPD scores of all 24 samples either passed the "CNV1" (red) and failed the "CNV1" with more than 100 bins dropped out (green). Red dash denotes the median MAPD scores at 0.18.

Table 4-2. Summary of all single cells analyzed for chromosomal copy numbers.

Sample_ID	Methods	Bincount.med	log2MAPD	dropout_-1	dropout_-2	n.segments	dropout1_filter	CNV4	CNV3	CNV1	CNV0
WGA4-3_7	WGA4	179	0.25	144	3	251	>100	NA	NA	NA	NA
WGA4-3_8	WGA4	180	0.25	147	13	262	>100	NA	NA	NA	NA
WGA4-3_9	WGA4	194	0.25	282	20	310	>100	NA	NA	NA	NA
WGA4-3_10	WGA4	155	0.26	106	12	162	>100	NA	NA	NA	NA
WGA4-3_11	WGA4	113	0.30	105	8	105	>100	NA	NA	NA	NA
WGA4-3_12	WGA4	355	0.18	9	0	30	<100	0	0	0	0
WGA4-3_13	WGA4	288	0.19	57	4	80	<100	0	0	5	0
WGA4-3_14	WGA4	322	0.19	26	0	55	<100	0	0	1	0
WGA4-3_15	WGA4	285	0.19	3	0	40	<100	0	0	0	0
WGA4-3_16	WGA4	309	0.16	6	0	36	<100	0	0	0	0
WGA4-3_17	WGA4	327	0.17	30	1	120	<100	0	0	6	0
WGA4-3_18	WGA4	405	0.22	62	2	93	<100	0	0	1	0
WGA4-3_19	WGA4	309	0.20	51	1	76	<100	0	0	2	0
WGA4-3_20	WGA4	365	0.19	19	0	80	<100	0	0	1	0
WGA4-3_21	WGA4	307	0.18	18	0	63	<100	0	0	3	0
WGA4-3_23	WGA4	302	0.18	7	0	45	<100	0	0	0	0
WGA4-3_24	WGA4	315	0.23	131	10	99	>100	NA	NA	NA	NA
WGA4-3_25	WGA4	316	0.15	0	0	27	<100	0	0	0	0
WGA4-3_26	WGA4	344	0.15	5	0	50	<100	0	0	0	0
WGA4-3_27	WGA4	319	0.17	3	0	29	<100	0	0	0	0
WGA4-3_28	WGA4	280	0.22	59	0	125	<100	0	0	4	0
WGA4-3_29	WGA4	246	0.15	2	0	52	<100	0	0	0	0
WGA4-3_31	WGA4	328	0.16	27	0	206	<100	0	0	3	0
WGA4-3_32	WGA4	298	0.14	17	0	97	<100	0	0	1	0
							Average	0	0	1.5	0
							Median	0	0	1	0

suggesting mapping artifact. The remaining 2 candidates were further validated by data reanalysis at 50,000 bins to test for binning artifacts and both of them no longer showed as copy number gain with 50,000-bin reanalysis, suggesting binning artifact. Therefore, there are no candidate CNV identified for copy number gains >2Mb. For copy number losses, 0 candidates were identified at CNV0 (i.e., homozygous deletion), and a total of 31 candidates were identified at CNV1 for heterozygous deletion across all samples (**Table 4-2**). Similarly, all 31 candidates were validated by BLAT and reanalysis at 50,000 bins to check for mapping and binning artifacts, respectively. All candidates were BLATed to mappable regions making mapping artifacts unlikely. 4 out 31 candidates no longer showed copy number loss in reanalysis with 50,000 bins. Among the remaining 27 candidates that were detected by both 60,000 and 50,000 bins, 19 of them appeared in chromosomes with additional segments of copy number losses, although some of these segments did not pass the arbitrary 2Mb minimal size requirement (**Figure 4-10A**). These candidates may be false positive, reflecting poorly amplified chromosomes, but could alternatively represent a recently characterized biological condition named chromothripsis, which involves gross chromosomal rearrangements and multiple copy number changes within a single chromosome. The other 8 candidates are isolated heterozygous copy number losses present in chromosomes that are otherwise euploid (**Figure 4-10B**); these candidates are more likely to be real though amplification dropout of specific loci cannot be ruled out.

From the previous study of somatic L1 insertion candidates, we learned that candidates that are shared by more than 1 cell are much more likely to be fully validated unless proven to be systematic artifact; one the other hand, all the false positive candidates are typically unique events. All 27 CNV candidates identified in this analysis are unique, suggesting that they are

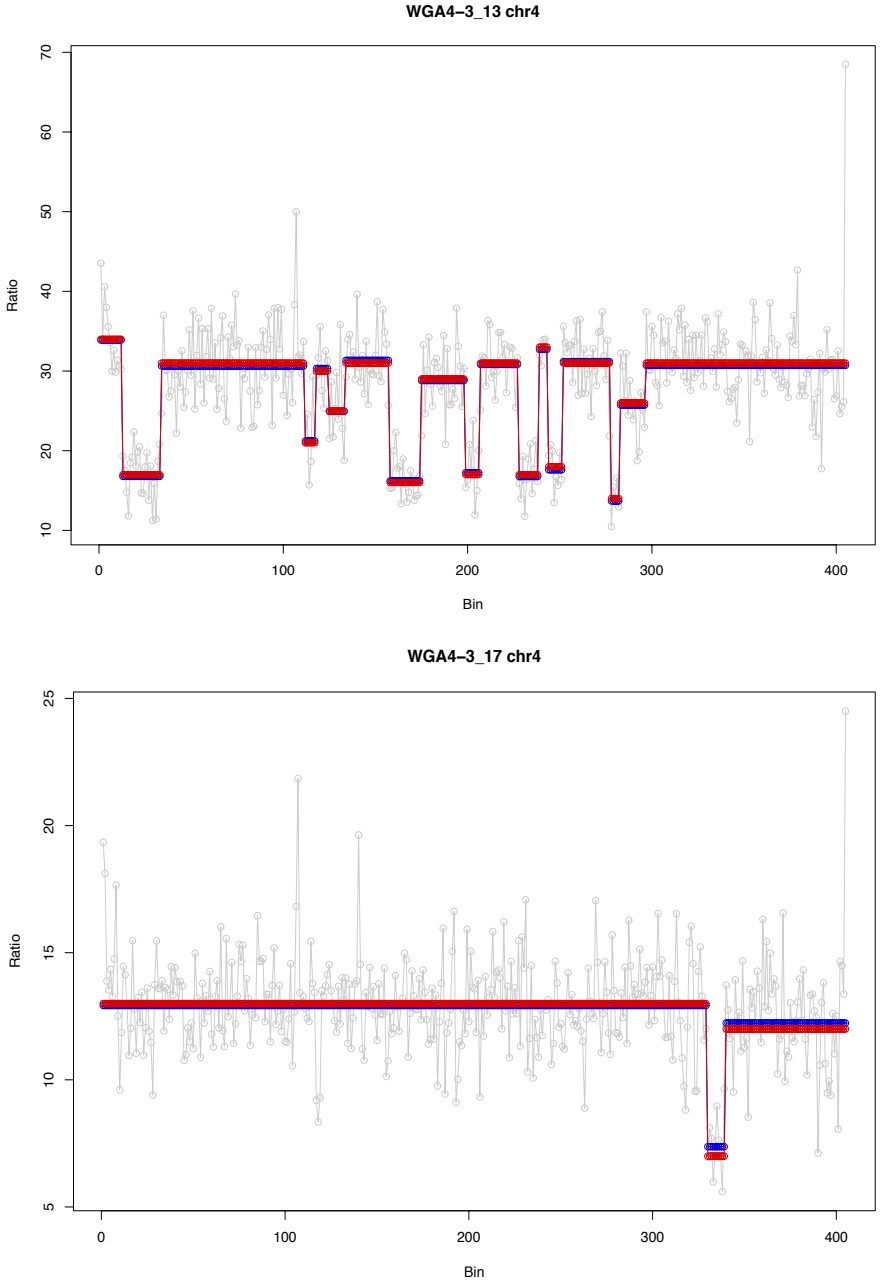


Figure 4-10. Representative chromosomes of with segmental CNV candidates.
(A) Chr4 from neuron WGA4_4643ctx_13. Multiple segmental copy number losses are observed within the same chromosome.

(B) Chr4 from neuron WGA4_4643ctx_17. A single segmental copy number loss is observed within the same chromosome.

X-axis represents the genome bins within each chromosome; Y-axis presents the copy number CN of each bin. Blue line denotes the segmented mean copy number and red line denotes the segmented copy number calls (in integers) of each chromosome. Calling confidence is generally higher when the blue approaches the red line.

either false positives or real somatic variants present at low percent mosaicism. Furthermore, we checked if any of the CNV candidates localize to one of the known *de novo* CNV "hot spots" from previous studies (Sanders et al. 2011) and we identified none. Collectively, our data showed that large somatic CNVs are not common at the single neuron level, with an average of 1.5 and a median of 1 CNV per neuron. More importantly, all the CNVs candidates identified are exclusively in the CNV1 state, which is known to be the most prominent type of amplification artifact (**Figure 4-3 and Figure 4-8**). Therefore, we expect the above estimated rate to be an overestimation given the difficulty in differentiating true variants from amplification dropout.

In general, copy number analysis of single cells at sub-chromosomal scale remains extremely challenging due to the technical limitations of amplification noise and locus dropout. Our analysis is unable to identify any convincing CNVs shared by multiple neurons or within the CNV "hotspot" regions previously identified down to 2Mb resolution from a total of 18 single neurons analyzed. This study sets an upper bound of large CNVs at copy number state 0, 3, and 4 to be no more than 6.6% at 95% confidence (0 events identified from 18 samples at 3 different states, CI = 0-6.6%) from the 18 samples analyzed. We expect the confidance interval to narrow, leading to a lower statistical upper bound when more samples analyzed in the future. We are unable to precisely define the rate of somatic CNV1s due to the inability of separating bona fide events from false positives, highlighting the limitation of current single cell amplification methods.

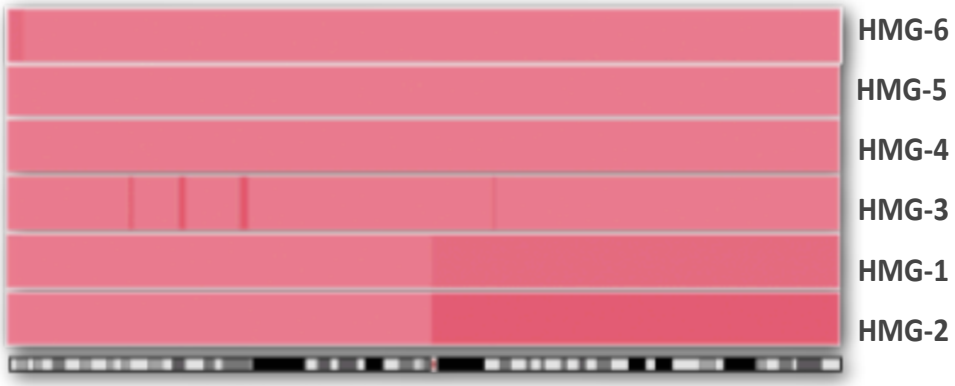
Chromosomal copy number analysis of single brain cells from hemimegalencephaly

In addition to the somatic copy number variants in normal human brains, we also studied the role of a somatic CNV in hemimegalencephaly, a brain overgrowth syndrome that is previously described to be caused by brain-specific somatic mutations (Poduri et al. 2012; Lee et

al. 2012). Two hemimegalencephaly cases were previously identified to have non-integer copy number increase of chromosome 1q, suggesting mosaic copy number gains in some but not all the cells from the affected tissue (**Figure 4-11A, B**). For one of the two cases affected brain tissue remained available (HMG-1). Copy number evaluation of SNP data showed increased signal for the entire q arm of chromosome 1 in the brain sample (**Figure 4-12A, B**), with an estimated copy number of 2.41 (SD 0.12). No other chromosomes displayed abnormal copy number (**Figure 4-12A**). Quantitative PCR (qPCR) confirmed the 1q copy number gain with the calculated copy number being 2.68 (SD 0.16), 2.76 (SD 0.20), and 2.73 (SD 0.13) at 1q21.3, 1q31.1, and 1q42.2, respectively (**Figure 4-12C**). High-resolution karyotype and qPCR of peripheral blood cells in the patient did not reveal any evidence of copy number increase of 1q in these nonbrain cells (**Figure 4-12C** and data not shown).

Single-cell copy number analysis was then performed on both the neuronal and non-neuronal populations from the affected brain tissue of HMG-1. Due to the poor tissue quality evident by the abnormal FACS scatter plot of HMG-1 (data not shown), neuronal versus non-neuronal 100-cell samples amplified by MDA were first analyzed for the presence of chromosome 1q copy number gains. An intermediate copy number gain of chromosome 1q was detected in both neuronal and non-neuronal cells (**Figure 4-13A**). The estimated copy number at chromosome 1q of neuronal and non-neuronal cells are 2.35 and 2.7, respectively. These estimates are in largely in agreement with previous estimates by SNP chip and qPCR (**Figure 4-12**), and are consistent with our similar study on somatic point mutation on *AKT3* (**Figure 1-3**), which affects both the neuronal and non-neuronal cells. The data suggest that the mutations occurred in a progenitor that gives rise to both neurons and glia. The higher chr1q copy number in non-neuronal cells suggests that a higher proportion of glial cells carrying the mutation

A



B

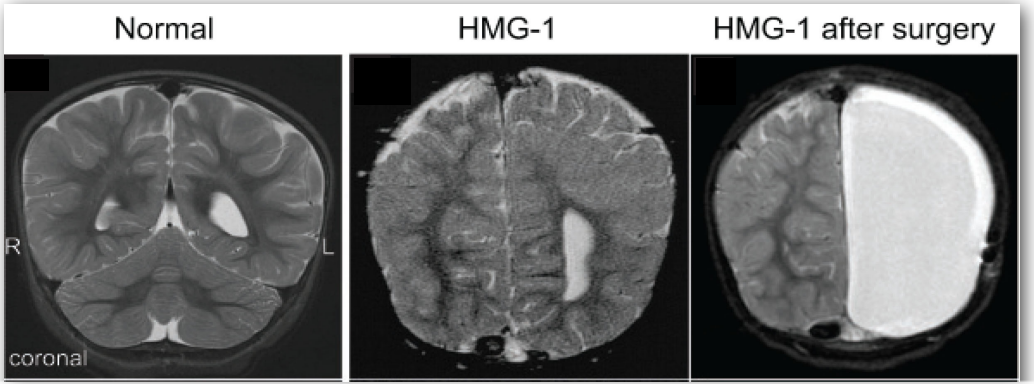


Figure 4-11. Copy number screening and MRI of hemimegalencephaly brains.
(A) Initial copy number analysis of Affymetrix 100K SNP data from 6 cases of HMG shows the trisomy of 1q in HMG-1 and HMG-2. The additional two cases were evaluated using Affymetrix 6.0 data (not shown). Dark pink indicates copy number 3, and light pink copy number 2 (normal). A diagram of chromosome 1 is presented (adapted from www.genome.ucsc.edu); note that there are no SNP probes in the centromeric regions.

(B) The first column shows an example of coronal T2-weighted and axial T2-weighted MRI images showing the brain of a normal 1 year old. Note the symmetric size of the right and left hemispheres, labeled R and L to denote standard MRI convention. The middle column shows a representative image of HMG-1. MRI before surgery showed left-sided hemispheric enlargement, abnormal cortical thickness and configuration, and enlarged left lateral ventricle in the coronal T2-weighted and axial T2-weighted images. The right hemisphere is smaller and appears normal. The right column shows representative MRI image after left hemispherectomy surgery, there is cerebrospinal fluid (CSF) where the abnormal hemisphere had been, seen as bright signal in coronal and axial images taken at approximately the same plane as the preoperative images.

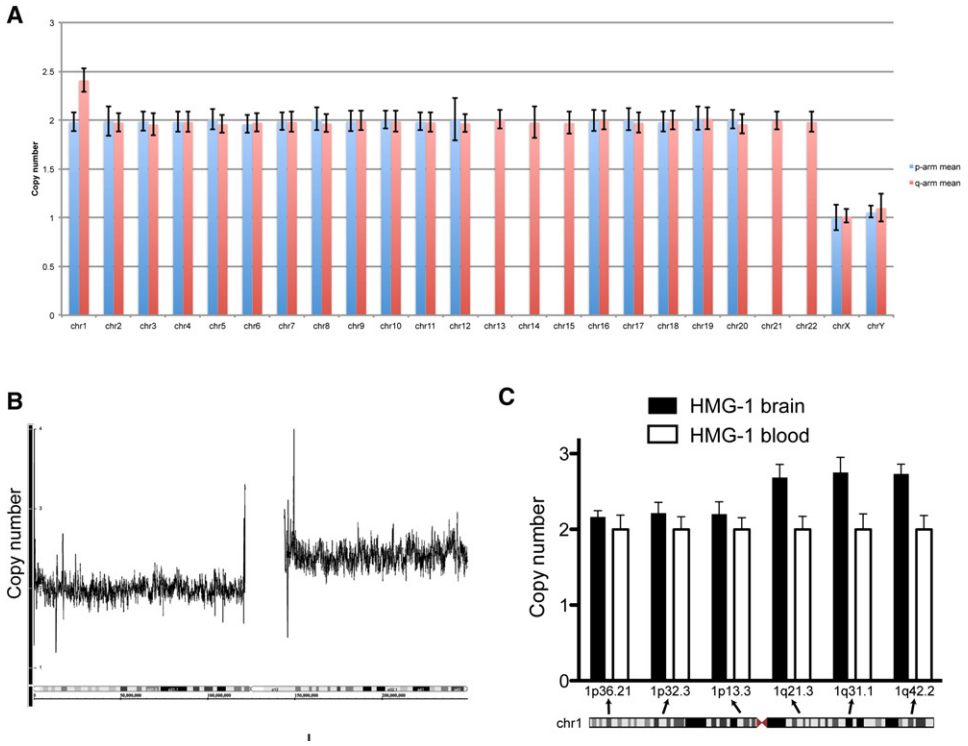


Figure 4-12. Mosaic copy number gain at chromosome 1q of HMG-1.
(A) Copy number for all of the chromosomes is shown for HMG-1; the estimated copy number for 1q is 2.41 (SD 0.12), consistent with mosaic trisomy 1q. Chromosome 1p, as well as the other autosomes, has normal copy number of 2, and chromosomes X and Y each show copy number of 1.

(B) Copy number evaluation of Affymetrix 6.0 data shows the gain in copy number at chromosome 1q for HMG-1, with the x axis representing nucleotide position along chromosome 1 and the y axis denoting copy number.

(C) Assuming a copy number of 2 for all regions in the DNA derived from leukocytes (white columns), the calculated copy number from the brain tissue (black columns) was 2.68 (SD 0.16) at 1q21.3, 2.76 (SD 0.20) at 1q31.1, and 2.73 (SD 0.13) at 1q42.2.

compared on neurons, perhaps reflecting their ability to continue to proliferate. The variable copy number estimates ranging from 2.41 to 2.76 may reflect the difference in the percent of cells carrying the mutation from different cortical regions sampled and/or the slight difference in neuronal versus glial composition of the sampled regions.

Chromosomal copy number analysis was then performed on single neurons isolated from the HMG-1 brain. Due to the poor tissue quality, which had gone through multiple freeze-thaw cycles before the analysis, most of the cells were amplified poorly and did not pass the MAPD QC threshold. Among all 45 single neurons sequenced, only 8 samples passed the threshold of MAPD <= 0.45; and 1 of 8 neurons was positive for chromosome 1q copy number gain (**Figure 4-13B**). Surprisingly, instead of detecting a trisomic chromosome 1q (CN=3) as previously assumed, tetrasomic chromosome 1q (CN=4) was detected from this single neuron, highlighting the importance of single cell copy number analysis in revealing the nature of mosaic copy number variations. Additional cells from both NeuN+ and NeuN- populations with MAPD scores that failed the QC threshold nonetheless independently confirmed the observation of tetrasomy 1q (**Figure 4-13C**). Tetrasomy 1q was previously described a mosaic state from embryos with nasopharyngeal teratomas, a rare neonatal neoplastic condition (Beverstock et al. 1999), providing further evidence for the pathogenic role of this CNV in driving over-proliferation. The previously identified tetrasomy 1q case exists as an isodicentric chromosome 1q (47, XX, +idic(1)(q10)), making it stably transmitted through cell division. Therefore, it is highly likely that the tetrasomy 1q cells from HMG-1 brain also represents an isodicentric 1q; interphase FISH analysis on the HMG-1 brain is to be carried out to confirm the finding.

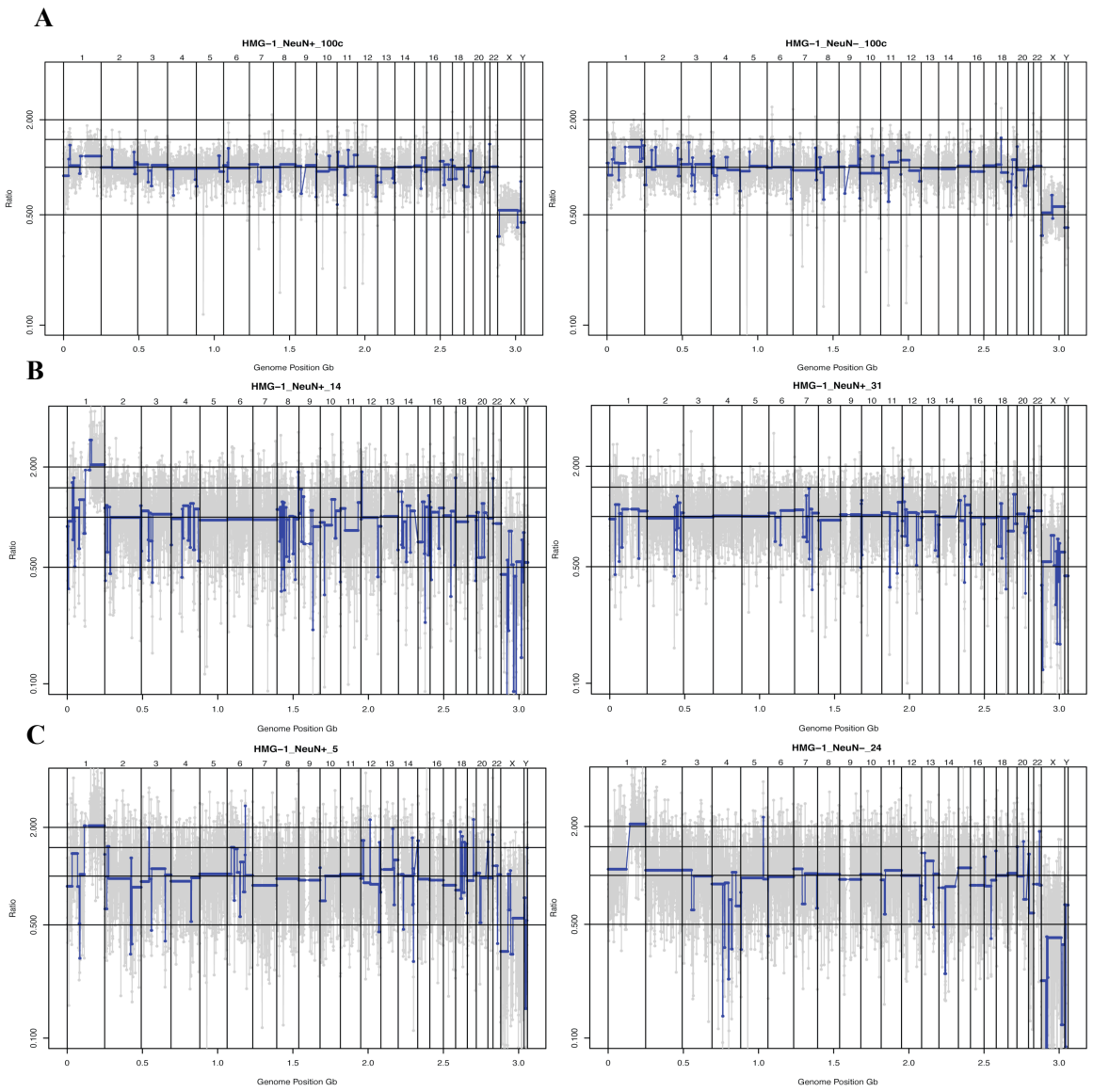


Figure 4-13. Single-cell copy number profiling of HMG-1.
(A) 100-cell neuronal versus non-neuronal cells from HMG-1 brain. Both show copy number increase at chromosome 1q.

(B) Representative single neurons from HMG-1 that passed the QC threshold. HMG-1_NeuN+_14 shows CN=4 at chromosome 1q; whereas HMG-1_NeuN+_31 shows CN=2 at chromosome 1q.

(C) Additional single cells that failed the QC threshold from both neuronal and non-neuronal population of HMG-1 brain but show tetrasomy 1q.

X-axis represents all 6,000 genomic bins arranger by their genomic positions by chromosomes. Log2 transformed Y-axis denotes the copy number ratio (CNR) respect to euploid genome.

Discussion

This chapter presents the first rigorous comparison of two independent single-cell whole genome amplification methods used for single cell copy number analysis. We conclude that WGA4 is a more reliable amplification method for single-cell copy number analysis, consistently producing amplified product with less stochastic amplification noise as well as less stochastic locus dropouts than MDA. We suspect that this advantage in amplification linearity is mostly due to the initial fragmentation of the genome into ~400kb fragments. Compared with the amplification fragments up to 50Mb in size by MDA, the 100-fold smaller fragment size allows more fragments to be independently amplified within any given genomic region, making the under or over amplification of fractions of the fragments less important to the overall copy number of the genomic region. This is evident by the increased stochastic noise and dropout by increased genomic bin sizes (**Figure 4-2**, and data not known). However, the gain of amplification linearity is at an expensive cost of genomic coverage since only ~10% of the initial fragments are successfully amplified, making the method unsuitable for studies of any other type of mutations (Navin et al. 2011; C. Zhang et al. 2013). Moreover, due to the pre-amplification fragmentation, WGA4 yields non-overlapping small fragments, which makes it incompatible with PCR based secondary validation. This is another lethal drawback of WGA4 since single-cell amplification is known to create a large number of technical false positives, secondary validations such as cloning of the breakpoints are crucial for making confident discoveries.

MDA poses significant advantages on the studies of a whole spectrum of somatic mutations types and is amendable for various secondary validations as discussed in

Chapter 3 and here we demonstrate that MDA can also be used for copy number analysis, at least at the chromosomal level. We achieved 100% sensitivity is detecting trisomy 18, which is 74Mb in size, from single cortical neurons. Trisomy 18 is one of the smallest chromosomes in the genome and copy number 3 is known to be most challenging to detect by any copy number algorithm. With such technical sensitivity, we can comfortably conclude that we will be able to detect most of the somatically aneuploid neurons if they are as common as previously suggested using other methods (Kingsbury et al. 2005).

From the 82 single cortical neurons, derived from 3 normal individuals, that we analyzed, we did not find any aneuploid neurons in which copy number alternations were limited to just one or two chromosomes, such as might result of chromosomal missegregation during mitosis due to multipolar spindles(Yang et al. 2003). In contrast, in the 3 out 82 neurons that showed chromosomal imbalances, they affected multiple chromosomes with copy number gains and losses at the whole chromosome or chromosome arm level, as well as alternating copy numbers within a single chromosome, a characteristic feature of chromosome "shredding" (**Figure 4-5, 6, 7**). These cells, if physiological, would not result from simple chromosomal missegregation, but rather would suggest more catastrophic events that produce global chromosomal rearrangements. No obvious mechanism is currently known to explain the observation. Chromotriphisis, a recently described phenomenon that leads to massive chromosomal breakage and rearrangements may be the most plausible biological explanation (Forment et al. 2012). However, the currently characterized chromotriphisis events are mostly restricted to a single chromosome and rarely affect multiple chromosomes at once

(Stephens et al. 2011). Although we cannot conclude definitely whether the observed aberrant cells are physiological, we successfully ruled out several obvious technical causes of the observations.

Regarding the 2 cells that showed large apparent homozygous deletions affecting multiple chromosomes (WGA4_4643ctx_30 and MDA_4638ctx_24), other groups have made similar observation. Baslan *et al.* (2012) who developed the first protocol of performing single-cell copy number analysis with WGA4 has described a similar observation and referred to it as "genome sector loss" (GSL), characterized by large homozygous deletions and patterns consistent with chromosomal "shredding". They observed ~5% of single cells with "genome sector loss" from a variety of sample sources, including cultured cells and postmortem samples from normal and malignant. Therefore, regardless its biological significance, this observation is not limited to the brain. It is also not a simple reflection of a specific whole-genome amplification method or lysis condition since similar observations have been made in both WGA4 and MDA amplified cells, as well as cells lysed by both proteinase K digestion (WGA4) and alkaline lysis (MDA and WGA4).

One potential technical cause of gross chromosome loss is loss of chromosomes during the nuclear sorting step, as both studies that report widespread chromosome loss used nucleus sorting as opposed to other single cell isolation methods. Nuclei are known to be fragile and thereby it is possible that a small fraction of the nuclei are shredded during the sorting process, leading to random loss of genomic segments. To test this, whole cells could be sorted or microfluidic devices used for cell/nuclei isolation for parallel analysis. It is also possible that the grossly abnormal cells are undergoing

apoptosis, although a ~5% frequency of apoptotic neurons in normal individuals seems to be too high. Postmortem apoptosis could also be a potential explanation; however, similar observation from cultured cells makes it less likely.

The other type of genome-wide chromosomal imbalance observed in our study, from a single neuron amplified by WGA4 (WGA4_4643ctx_22), does not exhibit homozygous genomic deletions, but instead shows chromosomal or sub-chromosomal copy number alternations covering four copy number states (CN=1, 2, 3, 4). The observation of some chromosomes with copy number > 2 makes it possible that a mitotic non-neuronal nucleus is accidently sorted or is attached to a neuronal nucleus. A recent study of single cell copy number profiling in S-phase cells reveal that DNA replication introduces (pseudo) false-positive copy number variations and the DNA-replication domain varies depending the stage of S-phase (Van der Aa et al. 2013). Alternatively, this cell can indeed be a partial tetrasomy cell that has undergone mitotic slippage followed by genome-wide chromosomal rearrangements. Such a result may also explain why earlier studies using interphase FISH technique detects cells with copy number ranging from 1-4 with probe against a single chromosome. In this case, the frequency of neurons with genome-wide chromosomal imbalance is estimated to be 1.2% (95% CI: 0.02%-6.6%).

Segmental CNV analysis of single cells is the most challenging aspect of single cell genomics, due to the lack of biological hallmarks for secondary validation. Our current results have suggested that CNVs at copy number 0, 3, or 4 larger than 2Mb are rare. However, due to the lack of positive controls, we cannot directly assess the sensitivity of our method at these 3 CNV states. The lack of positive controls reflects the

fact that CNVs at this large size barely occur in the germline of healthy individuals and none of the 3 normal individuals analyzed have germline CNVs larger than 2Mb that could be used as positive control. Alternatively, disease-causing, large CNVs can be used as positive controls in the future to assess the sensitivity and limitation of the method.

An average of 1.5 events/cell of heterozygous deletion (i.e. CNV1) larger than 2Mb was detected. These candidates should be further validated by an LOH test, which is not possible with the current data due to the ultra low sequencing depth. More than half of the heterozygous deletions are localized to chromosomes with additional deletions, suggestive of either poor amplification of the whole chromosome or chromothripsis (**Appendix Table 4-1**). Further higher coverage sequencing and LOH analysis could differentiate the two possibilities. Nevertheless, this CNV frequency is likely to be a significant over-estimation of the true frequency since some CNV calls probably represent false-positive dropouts. The true specificity is impossible to assess with the current method due to the lack of secondary validation that can be applied independent of the amplification bias and the lack of gold standard negative controls since the genome of every single cells can be different.

The identification of somatic mosaic tetrasomy 1q, which likely represents an isodicentric 1q, and that drives the over-proliferation of neural progenitors to cause hemimegalencephaly is an important finding in epilepsy genetics. It points to a second molecular mechanism, besides the PI3K-AKT-mTOR activating mutations, that can cause this non-malignant overgrowth syndrome. More interestingly, both of the PI3KC2B and AKT3 genes are localized to chromosome 1q; therefore, it's tempting to hypothesize that the dosage increase of these two genes is sufficient to hyper-activate the pathway to

provide proliferative advantages is the affected cells. It also likely explains why somatic chromosome 1q copy number gains are highly enriched in cancer cells (Beroukhim et al. 2010). Although aneuploidy are often regarded as anti-proliferative, here we provided an exceptional example of aneuploid cells that show pathogenic proliferative advantages, providing insights into the ongoing debate about whether aneuploidy can be a causal driver of cancers (Siegel & Amon 2012).

Materials and Methods

Copy number analysis pipeline

The copy number analysis pipeline is adapted but modified from Baslan, et al. (2012). The data analysis flowchart is presented in **Figure 4-14**.To create bin boundaries, 20 million 50bp simulation reads against hg19 reference and mapped with bowtie -v 2 -m 1 --best --strata settings (Langmead et al. 2009). Boundaries of 6,000, 20,000, and 50,000 genomes bins are generated with variable sizes to ensure equal number of reads inside each bins. The average bin size of 6,000, 20,000 and 50,000 are 500kb, 150kb and 60kb, respectively. Generating equal-read instead of equal-size genomic bins controls for the mappability of the genome so that each bin receives a comparable number of reads and thereby a similar data variance from random sampling. The percent of GC-content of each calculated bin was then measured for later GC normalization. All sequence reads are first demultiplexed by CASAVA allowing for 1 mismatch in the 6-bp index sequences. Sequence reads are further trimmed based on the FASTQC score allowing at least 38bp reads for bowtie mapping with the same setting. Mapped stats are reviewed to samples with at least 80% uniquely mapped reads for MDA and 40% uniquely mapped reads for

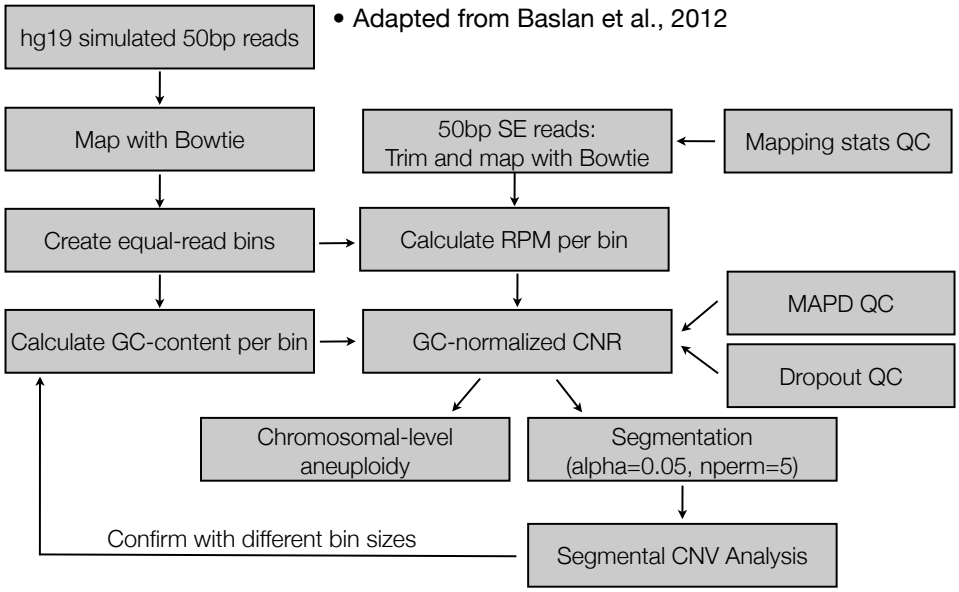


Figure 4-14. Illustration of the scCNV data analysis pipeline.

WGA4 are allowed to proceed forward. The low mapping percentage of WGA4 samples is due to the additional adaptors added to the amplicons. Then the number of reads in each bin is counted (RPB_i) and normalized by the total number of reads of the particular sample ($RPMB_i$).

The copy number ratio (CNR) is then calculated by: $CNR_i = RPB_i/average(RPB_i)$. Chromosomal copy number is then calculated as previously described in Chapter 3. Any chromosome with a copy number (CN) > 1.5 or <0.5 is called as copy number gain and loss, respectively. In parallel, the GC-normalized CNR (GCNR) (see detailed in Chapter 3) is calculated for each bin, and the data set is subject to segmentation with parameters alpha=0.05 and number of permutation (nperm)=5 using the DNAcopy package (Venkatraman & Olshen 2007). A quality control step on amplification noise measured by MAPD score is applied to all the samples passed the mapping stats QC (see details in next section). Samples with MAPD score <= 0.45 are allowed to proceed to chromosomal copy number analysis and segmentation. We found a strong correlation between chromosomal copy number alternations and high MAPD scores, suggesting that poor quality samples with high MAPD scores tend to produce false positive results and thereby should be excluded from further analysis (**Figure 4-4**). After segmentation, the copy number profiles of each sample are plotted by the GCNR value in gray and segmented GCNR (seg.GCNR) value in blue of each genomic bin across the genome. The aneuploidy calls on a chromosomal level are further validated with the segmentation. Again, all trisomy 18 cells that passed the QC threshold showed copy number gain at chromosome 18 after segmentation. The three cells with chromosomal imbalance are further analyzed at every single chromosome to reveal the pattern of copy number

alternations. For segmental CNV analysis, only WGA4 amplified single neurons are used. An additional quality-control step was incorporated by counting the number of bins with a copy number loss at 1 or more (log2GCNR <= - 1) from each sample. A threshold of 100 bins/genome with copy number <= 1 is used to eliminate poor quality cells with increased technical dropouts from further analysis. Segmental CNVs with size larger than 2Mb are called of the segmentation data based on the following criteria: CNV0 when seg.log2GCNR < -1.5; CNV1 when -1.5 <= seg.log2GCNR < -0.6; CNV2 when -0.6 <= seg.log2GCNR <= 0.4; CNV3 when 0.4 < seg.log2GCNR <= 0.8 and CNV4 when seg.log2GCNR > 0.8. All the CNV candidates were first called with 6,000 bins and then validated by rerunning the pipeline at 50,000 bins; only the candidates that are called by both analyses are retained as the final candidates. In parallel, all the candidates are BLATed against the UCSC hg19 reference genome and some of the candidate regions are turned out to be unmappable regions such as centromeres and telomeres. These candidates are also excluded from the final list as they are caused mapping and binning artifacts.

MAPD QC metrics

The MAPD QC metric was developed by adapting the Affymetrix multiple absolute pairwise differences algorithm (Affymetrix, Inc. 2008). MAPD = median($|\log2GCNR_{i+1} - \log2GCNR_i|$) where i stands for individual bins. A MAPD threshold of 0.45 is used for all single cell genome amplified samples because it appears to be the border line where samples with noisy copy number profiles start to cause false positive aneuploidy calls (**Figure 4-4**).

Copy number screen, SNP-chip and qPCRs

We obtained eight samples of flash-frozen brain tissue resected during focal epilepsy surgery for HMG. DNA was extracted by using standard methods and was then digested, amplified, and hybridized to Affymetrix 100K SNP arrays for six of the samples (Affymetrix). In the original arrays (e.g., 100K), copy number was assessed based on intensity of signal from each SNP. For the Affymetrix 6.0 arrays, copy number probes are included in addition to the full array of SNPs, and both are used for quantitation. The Gaussian- smoothed signal log2-ratio of all probe intensities normalized to a reference of 270 normal HapMap samples was calculated by Affymetrix Genotyping Console with standard settings. Additional DNA from HMG-1 and two other samples was assessed by using the Affymetrix 6.0 SNP array. The software dChipSNP was used for analysis.

For HMG-1 and HMG-2, we performed qPCR in cases in which copy number change was detected. Primers were designed to 1q44 and 1p21.1. DNA from two control individuals (Promega) was used for comparison. We repeated qPCR in an additional specimen from HMG-1 for confirmation by using primers targeting 1p (1p13.3, 1p32.3, and 1p36.2) and 1q (1q21.3, 1q31.1, and 1q42.2).

Leukocytes were obtained from six of the cases; DNA was extracted by using standard methods and was used for SNP analysis as above. For HMG-1, we performed SNP analysis and clinical karyotype to assess for the presence of mosaic copy number gain at 1q in peripheral blood leukocytes (evaluating 50 cells to detect even a low level of mosaicism).

Reference

Affymetrix, Inc., 2008. Whitepaper, Median of the Absolute Values of all Pairwise Differences and Quality Control on Affymetrix Genome-Wide Human SNP Array 6.0. pp.1–8.

Baslan, T. et al., 2012. Genome-wide copy number analysis of single cells. 7(6), pp.1024–1041.

Beroukhim, R. et al., 2010. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283), pp.899–905.

Beverstock, G.C. et al., 1999. Nasopharyngeal teratoma and mosaic tetrasomy 1q detected at amniocentesis. A case report and review of the literature. *Cancer genetics and cytogenetics*, 115(1), pp.11–18.

Cheng, J. et al., 2011. Single-cell copy number variation detection. *Genome biology*, 12(8), p.R80.

Faggioli, F., Vijg, J. & Montagna, C., 2011. Chromosomal aneuploidy in the aging brain. *Mechanisms of Ageing and Development*, 132(8-9), pp.429–436.

Forment, J.V., Kaidi, A. & Jackson, S.P., 2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nature Reviews Cancer*, 12(10), pp.663–670.

Jacobs, K.B. et al., 2012. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics*, 44(6), pp.651–658.

Kingsbury, M.A. et al., 2005. Aneuploid neurons are functionally active and integrated into brain circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17), pp.6143–6147.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Laurie, C.C. et al., 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, 44(6), pp.642–650.

Lee, J.H. et al., 2012. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nature Genetics*, 44(8), pp.941–945.

Navin, N. et al., 2011. Tumour evolution inferred by single-cell sequencing. *Nature Genetics*, 472(7341), pp.90–94.

Olshen, A.B. et al., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*, 5(4), pp.557–572.

O'Huallachain, M. et al., 2012. Extensive genetic variation in somatic human tissues. *Proceedings of the National Academy of Sciences*, 109(44), pp.18018–18023.

O'Huallachain, M., Weissman, S.M. & Snyder, M.P., 2013. The variable somatic genome. *Cell Cycle*, 12(1), pp.5–6.

Poduri, A. et al., 2012. Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron*, 74(1), pp.41–48.

Sanders, S.J. et al., 2011. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, 70(5), pp.863–885.

Siegel, J.J. & Amon, A., 2012. New insights into the troubles of aneuploidy. *Annual review of cell and developmental biology*, 28, pp.189–214.

Stephens, P.J. et al., 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1), pp.27–40.

Van der Aa, N. et al., 2013. Genome-wide copy number profiling of single cells in S-phase reveals DNA-replication domains. *Nucleic acids research*.

van Echten-Arends, J. et al., 2011. Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Human reproduction update*, 17(5), pp.620–627.

Vanneste, E. et al., 2009. Chromosome instability is common in human cleavage-stage embryos. *Nature medicine*, 15(5), pp.577–583.

Venkatraman, E.S. & Olshen, A.B., 2007. DNAcopy: A Package for analyzing DNA copy data. *Department of Epidemiology and Biostatistics. Memorial Sloan-Kettering Cancer Center*.

Wang, J. et al., 2012. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell*, 150(2), pp.402–412.

Yang, A.H. et al., 2003. Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(32), pp.10454–10462.

Yin, X. et al., 2013. Massively Parallel Sequencing for Chromosomal Abnormality Testing in Trophectoderm Cells of Human Blastocysts. *Biology of reproduction*.

Zhang, C. et al., 2013. A Single Cell Level Based Method for Copy Number Variation Analysis by Low Coverage Massively Parallel Sequencing. *PloS one*, 8(1), p.e54236.

Zhang, F. et al., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10, pp.451–481.

Zong, C. et al., 2012. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science (New York, N.Y.)*, 338(6114), pp.1622–1626.

Chapter 5: Conclusions and Future Directions

Genetic Diversity and Diseases

This dissertation, along with the dissertation of Gilad Evrony, serves as a proof-of-principle study on how to systematically identify and quality genetic variabilities within a human brain, and it provides a foundation for future work on study of functional consequences of various types of somatic mutational events during neurodevelopment and aging. In this study, we show that the individual genomes of single neurons are indeed variable, with low frequency of somatic mutations acquired during the process of neurogenesis. We demonstrate that both L1 retrotransposition and aneuploidy—the two mutation types that were proposed to have increased prevalence in the human brain cells—are in fact rare with little evidence of brain specific elevation. Therefore, they are unlikely to serve as an obligated generator of genetic diversity within the human brain. However, our work presents only an incomplete survey of a subset of somatic mutations. Therefore, it is still possible that other types of somatic variants could be actively generated during human brain development, to diversify the genome of brain cells. To date, the only known mechanism for such genetic diversification is the V(D)J mechanism, which functions exclusively in the immune system (Jung et al. 2006). However, there is no clear evidence suggesting a similar mechanism active in the CNS; therefore, a CNS-specific increase in genomic diversity seems unlikely based on current knowledge.

On the other hand, neurons are the post-mitotic cells that live for decades. They are expected to accumulate DNA damages caused by endogenous (e.g. ROS) and environmental insults throughout individual's lifespan. During this process, additional somatic mutations continue to be accumulated. Their functional consequences are unlikely to be obvious in the early stages of individual's lifetime; however, these variants

and damages are accumulative, so they would become more and more likely to alter neuronal function with aging. Such a mechanism coincides with the development of neurodegenerative diseases with aging. Although genome-wide assessment of increased genetic variability during aging has not been available, studies on specific disease-relevant loci have provided ample evidence on the ongoing somatic mutagenesis in human brains. For example, the length of the CAG repeats causing Huntington's disease (HD) continues to expand in postmitotic neurons. Further evidence showed that such an ongoing mutagenesis are regulated in a regional-specific fashion with neurons from striatum harboring the highest genomic instability (Gonitel et al. 2008). Performing single-neuron whole-genome sequencing from young versus aged brains, one can immediately start asking questions such that whether there is any type of somatic mutations overrepresented during the aging process and whether there is any particular genomic regions ("hotspots") susceptible for a given type of somatic variant. More importantly, the functional consequences of these potential genetic variants can be studied to definitively study the correlation between genomic instability and neurodegenerative diseases. Single-cell sequencing technique has become a valuable tool to interrogate these questions as each of the somatic variants occurred post-mitotically are unique to the individual neuron and hence cannot be easily identified by bulk sequencing.

Aneuploidy

The accumulation of aneuploid neurons during neurogenesis as well as during the aging process has been proposed as one of the major contributors to genetic variations within the human brain. Additionally, PGD single-cell screens on cleavage stage

blastomeres used for IVF procedures have revealed an unexpected high frequency of aneuploid cells and mosaic aneuploid embryos, ranging from 30%-80% based on a number of studies (van Echten-Arends et al. 2011; Vanneste et al. 2009; Yin et al. 2013). Assuming this is not an IVF-specific artifact, it suggests that many of us are developed from mosaic embryos. Although it is known that most aneuploid cells—with a few exceptions including trisomy and tetrasomy 1q—confer proliferative disadvantages; and therefore, they are likely to be outcompeted during embryonic development. It is still tempting to ask to which degree does our normal body can tolerate aneuploidy mosaicism and whether this is determined at a tissue/organ-specific manner. Two organs have been shown to harbor high levels of tissue-specific aneuploidy: one of which is the brain. Previous studies reported up to 30% aneuploidy within mouse neural progenitors and up to 10% within matured neurons from mouse and human adult brains (Rehen et al. 2001; Rehen et al. 2005). The hepatocyte in liver is the other cell type that has been observed to have high frequencies of aneuploidy; and the frequency appears to increase during aging and in response to toxic stresses (Duncan et al. 2010; Duncan et al. 2012). Despite the overall reduction of fitness to aneuploid cells, acquired aneuploidy has been proposed as a fast adaptive mechanism to external stresses under short-term selective pressure (Tang & Amon 2013). Conceivably, the very early stages of embryonic development could be highly stressful and therefore results in an extremely high rate of aneuploidy.

In the context of brain development and aging brains, a high frequency of aneuploid cells among the neural progenitors presumably reflects the replicative-related stress during neurogenesis. On the other hand, those aneuploid progenitors are expected to be less proliferative compared to their wildtype competitors; therefore, they are more

likely to be outcompeted, or alternatively, terminally differentiate into neurons and survive. If the latter case were true, we would expect to see high prevalance of aneuploid post-mitotic neurons as some of the previous studies suggested (Rehen et al. 2005; Westra et al. 2008; Yurov et al. 2007). However, based on our comprehensive genome-wide chromosomal copy number profiling of 82 cortical neurons from 3 normal brains (at age of 17, 15, and 42), we are unable to identify a single neuron with discrete chromosomal copy number alternation limited to one or two chromosome, as predicted outcome of mitotic missegregation. Instead, we identified three neurons with genome-wide chromosomal imbalances. Two out of the three neurons harbor large homozygous deletions, a phenomenon independently described by Baslan et al. (2012). Since both studies used nucleus sorting to isolate single cells, we suspect that these two single neurons with large homozygous deletions are probably technical artifacts associated with nucleus sorting. The one additional aberrant neuron harbors multiple chromosomal gains and losses with 4 distinct copy number states (CNV1, 2, 3, 4). It can either be a mitotic cell falsely sorted or a partial tetraploid neuron that had gone through major catastrophic events, leading to unbalanced genome-wide chromosomal rearrangements. In any case, the frequency of such aberrant cells remains low (1/82); and therefore they are unlikely pose functional impact on the development of normal human brains. Nevertheless, it would be interesting to follow up on the molecular mechanism of grossly aneuploid neurons. One potential mechanism, analogous to the polyploidy/aneuploid hepatocytes, predicts the aneuploid neuron as a result of a progenitor cell undergone mitotic slippage followed by random multipolar division (Tang & Amon 2013). Alternatively, the grossly altered genome can be a secondary effect of misregulated cell division in the first place

(Ganem & Pellman 2012; Crasta et al. 2012). It is also interesting to note that multiple previous studies have observed an increase of aneuploid brain cells during aging (Yurov et al. 2007; Iourov et al. 2009; Kingsbury et al. 2006). Although the absolute quantification of aneuploid rates based on the interphase FISH studies can be indirect and inaccurate, the consistent relative increase of aneuploidy frequency in aged brains is informative. These observations together all suggest that increased chemical stress level (e.g. ROS) during aging of the human brain can lead to the accumulation of aneuploid cells. Further single cell copy number analysis on aged brain (age >70) could test whether the frequency of grossly aneuploid cells (analyzed at the age at 42) increases in an older brain. And if this is the case, it coincides with the previous observation shown by interphase FISH, and it suggests that the "aneuploid" cells detected by interphase FISH are instead more likely to be cells with unbalanced genome-wide chromosome rearrangements. The last piece of the puzzle is that how would a post-mitotic neuron turn into an aneuploid cells without going through further cell division? Further mechanistic insight into the formation of the cells with genome-wide chromosomal imbalances would be required to answer the question.

References

Baslan, T. et al., 2012. Genome-wide copy number analysis of single cells. 7(6), pp.1024–1041.

Crasta, K. et al., 2012. DNA breaks and chromosome pulverization from errors in mitosis. *Nature*, 482(7383), pp.53–58.

Duncan, A.W. et al., 2012. Aneuploidy as a mechanism for stress-induced liver adaptation. *The Journal of clinical investigation*, 122(9), pp.3307–3315.

Duncan, A.W. et al., 2010. The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature*, 467(7316), pp.707–710.

Ganem, N.J. & Pellman, D., 2012. Linking abnormal mitosis to the acquisition of DNA damage. *The Journal of cell biology*, 199(6), pp.871–881.

Gonitel, R. et al., 2008. DNA instability in postmitotic neurons. *Proceedings of the National Academy of Sciences*, 105(9), pp.3467–3472.

Iourov, I.Y. et al., 2009. Aneuploidy in the normal, Alzheimer's disease and ataxia-telangiectasia brain: differential expression and pathological meaning. *Neurobiology of disease*, 34(2), pp.212–220.

Jung, D. et al., 2006. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology*, 24, pp.541–570.

Kingsbury, M.A. et al., 2006. Aneuploidy in the normal and diseased brain. *Cellular and molecular life sciences : CMLS*, 63(22), pp.2626–2641.

Rehen, S.K. et al., 2001. Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), pp.13361–13366.

Rehen, S.K. et al., 2005. Constitutional aneuploidy in the normal human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(9), pp.2176–2180.

Tang, Y.-C. & Amon, A., 2013. Gene copy-number alterations: a cost-benefit analysis. *Cell*, 152(3), pp.394–405.

van Echten-Arends, J. et al., 2011. Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Human reproduction update*, 17(5), pp.620–627.

Vanneste, E. et al., 2009. Chromosome instability is common in human cleavage-stage embryos. *Nature medicine*, 15(5), pp.577–583.

Westra, J.W. et al., 2008. Aneuploid mosaicism in the developing and adult cerebellar cortex. *The Journal of comparative neurology*, 507(6), pp.1944–1951.

Yin, X. et al., 2013. Massively Parallel Sequencing for Chromosomal Abnormality Testing in Trophectoderm Cells of Human Blastocysts. *Biology of reproduction*.

Yurov, Y.B. et al., 2007. Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PloS one*, 2(6), p.e558.

Appendix Table 4-1. Summary of segmental CNVs identified from WGA4 amplified single neurons.

ID	Chrom	Start	End	Num. bin	Size (MB)	Num.samples shared	Copy number	Comment
CNV1_1	4	95514470	98688435	8	3.17	1 (13)	1	multiple dropouts
CNV1_2	4	108959992	113614987	11	4.65	1 (13)	1	multiple dropouts
CNV1_3	4	116309137	119638575	8	3.33	1 (13)	1	multiple dropouts
CNV1_5	17	9022233	11365374	6	2.34	1 (13)	1	multiple dropouts
CNV1_6	17	56463429	58563209	5	2.10	1 (13)	1	multiple dropouts
CNV1_7	20	43683943	45958956	6	2.28	1 (14)	1	look real, remains in 50k
CNV1_8	2	82924718	85207231	6	2.28	1 (17)	1	multiple dropouts
CNV1_9	3	10144745	12434756	6	2.29	1 (17)	1	multiple dropouts
CNV1_10	4	156172930	160256493	10	4.08	1 (17)	1	look real, remains in 50k
CNV1_11	5	112878401	115198557	6	2.32	1 (17)	1	multiple dropouts
CNV1_12	7	150460160	154011942	8	3.55	1 (17)	1	look real, remains in 50k
CNV1_13	10	130418442	132609109	6	2.19	1 (17)	1	multiple dropouts
CNV1_15	18	7880279	10169561	6	2.29	1 (18)	1	look real, remains in 50k
CNV1_16	14	99679698	102844929	8	3.17	1 (19)	1	look real, remains in 50k
CNV1_17	15	47905237	52033465	10	4.13	1 (19)	1	multiple dropouts
CNV1_18	5	33629556	37072917	8	3.44	1 (20)	1	look real, remains in 50k
CNV1_20	8	14091831	16323312	6	2.23	1 (21)	1	look real, remains in 50k
CNV1_21	14	64196851	66531916	6	2.34	1 (21)	1	look real, remains in 50k
CNV1_22	19	20152682	22376604	5	2.22	1 (21)	1	multiple dropouts
CNV1_23	2	64143115	66916655	7	2.77	1 (28)	1	multiple dropouts
CNV1_25	6	103749166	106972538	8	3.22	1 (28)	1	multiple dropouts
CNV1_26	11	7990609	11273782	8	3.28	1 (28)	1	multiple dropouts
CNV1_27	14	55414198	58198919	7	2.78	1 (28)	1	multiple dropouts
CNV1_28	3	4637443	9639508	12	5.00	1 (31)	1	multiple dropouts
CNV1_29	3	122844551	125094833	6	2.25	1 (31)	1	multiple dropouts
CNV1_30	10	118456139	120693419	6	2.24	1 (31)	1	multiple dropouts
CNV1_31	23	49039875	53019359	7	3.98	1 (32)	1	look real, remains in 50k